
ERIK J. LARSON

EL MITO DE LA INTELIGENCIA ARTIFICIAL

Por qué las máquinas no pueden pensar como nosotros lo hacemos

Shackleton books


ERIK J. LARSON

EL MITO DE LA INTELIGENCIA ARTIFICIAL

Por qué las máquinas no pueden pensar como nosotros lo hacemos

Shackleton books

EL MITO DE LA INTELIGENCIA ARTIFICIAL

EL MITO DE LA INTELIGENCIA ARTIFICIAL

Por qué las máquinas no pueden pensar como nosotros lo hacemos

ERIK J. LARSON

Traducción de Milo J. Krmpotić

Shackleton

—books—

El mito de la Inteligencia Artificial. Por qué las máquinas no pueden pensar como nosotros lo hacemos

Título original: The Myth of Artificial Intelligence. Why computers can't think the way we do

La presente edición se publica en acuerdo con Harvard University Press a través de International Editors' Co

Shackleton

—books—

@Shackletonbooks

www.shackletonbooks.com

Realización editorial: La Letra, S. L.

Diseño de cubierta: Pau Taverna

Conversión a ebook: Iglú ebooks

ISBN: 978-84-1361-202-7

Índice

Para Brooke y Ben

Introducción

En las páginas de este libro vas a leer acerca del mito de la inteligencia artificial. Lo de «mito» no se refiere a la imposibilidad de una IA verdadera. A ese respecto, el futuro de la IA es un misterio para la ciencia. El mito de la inteligencia artificial consiste en afirmar que su llegada es inevitable, mera cuestión de tiempo —que nos hemos adentrado ya en el sendero que conducirá a una IA de nivel humano, y más tarde a una superinteligencia—. No es así. Ese sendero existe solo en nuestra imaginación. Sin embargo, el carácter inevitable de la IA se encuentra tan arraigado en el debate popular —promovido por los expertos de los medios de comunicación, por referentes intelectuales como Elon Musk e incluso por numerosos científicos de IA (aunque desde luego no por todos ellos)— que, a menudo, cualquier pega que se le ponga se considera una forma de ludismo, o por lo menos una visión corta de miras sobre el futuro de la tecnología y un fracaso peligroso a la hora de prepararse para un mundo de máquinas inteligentes.

Tal y como os voy a mostrar, la ciencia de la IA ha revelado un misterio de grandes dimensiones en el núcleo de la inteligencia, y en la actualidad nadie tiene la menor idea de cómo resolverlo. Los partidarios de la IA cuentan con inmensos incentivos para minimizar sus limitaciones. Al fin y al cabo, la IA es un negocio enorme y tiene una presencia cada vez más predominante en la cultura. No obstante, nos guste o no, la posibilidad de un futuro de sistemas de IA se encuentra limitada por lo que sabemos en la actualidad sobre la naturaleza de la inteligencia. Y aquí deberíamos afirmarlo con franqueza: todas las pruebas sugieren que las inteligencias humana y artificial son radicalmente diferentes. El mito de la IA insiste en que esas diferencias son solo temporales, y en que la aparición de sistemas más potentes acabará por erradicarlas. Futurólogos como Ray Kurzweil y el filósofo Nick Bostrom, prominentes proveedores del mito, hablan no solo como si la IA de nivel humano resultara inevitable, sino como si, al poco de su llegada, las máquinas superinteligentes fueran a dejarnos muy atrás.

Este libro explica dos aspectos importantes del mito de la IA, uno de tipo científico y otro cultural. La parte científica del mito asume que solo tenemos que seguir «desnudando la cebolla» del desafío de la inteligencia general,

avanzando en hitos restrictivos de la inteligencia como la participación en juegos o el reconocimiento de imágenes. Se trata de un error grave: el éxito en las aplicaciones débiles no nos acerca ni un solo paso a la inteligencia general. Las inferencias que requieren los sistemas de cara a alcanzar una inteligencia general —leer el periódico, o mantener una conversación elemental, o ejercer de ayudante, como el robot Rosie de Los Supersónicos— no se pueden programar, aprender ni diseñar a partir de nuestro conocimiento actual de la IA. Al aplicar con éxito versiones de inteligencia más simples y débiles, que se benefician del uso de ordenadores más rápidos y de montones de datos, no estamos obteniendo un avance progresivo, sino que nos limitamos a recoger sus frutos maduros. El salto hacia un «sentido común» general es completamente diferente, y no se conoce camino alguno que lleve de lo uno a lo otro. No existe ningún algoritmo para la inteligencia general. Y tenemos buenos motivos para mostrarnos escépticos ante la idea de que dicho algoritmo vaya a surgir de nuevas tentativas con los sistemas de aprendizaje profundo o de cualquier otra aproximación popular en la actualidad. Resulta mucho más probable que vaya a requerir de un avance científico de primer orden, y ahora mismo nadie tiene la más remota idea del aspecto que tendría ese avance, y mucho menos de los detalles que conducirán a él.

La mitología sobre la IA es negativa, pues, porque oculta un misterio científico bajo la cháchara interminable del progreso continuado. El mito sostiene la creencia en un éxito inevitable, pero el respeto genuino por la ciencia debería hacer que volviéramos a la casilla de salida. Eso nos conduce al segundo tema de estas páginas: las consecuencias culturales del mito. Perseguir un mito no es la mejor manera de obtener «inversiones expertas», y ni siquiera una posición neutral. Es malo para la ciencia y es malo para nosotros. ¿Por qué? Un motivo es que resulta poco probable que alcancemos innovaciones si decidimos ignorar un misterio tan básico en vez de afrontarlo. La versión saludable de la cultura de la innovación pone el énfasis en la exploración de lo que se desconoce, no en dar bombo a la ampliación de unos métodos ya existentes —sobre todo cuando esos métodos se han revelado inadecuados para llevarnos mucho más allá—. La mitología acerca del éxito inevitable de la IA tiende a extinguir la cultura misma de la invención, tan necesaria para obtener un avance real —con la IA de nivel humano o sin ella—. El mito también fomenta la resignación ante el progresivo avance hacia una tierra de máquinas, donde la invención genuina se deja de lado en favor de charlas futuristas que defienden los métodos actuales, a menudo desde intereses particulares.

¿Quién debería leer este libro? Sin duda, cualquier persona que se emocione con la idea de la IA pero que se esté preguntando por qué siempre aparece a diez o veinte años vista. Hay un motivo científico para ello, que explicaré. También deberías leer este libro si piensas que el progreso de la IA hacia la superinteligencia es inevitable y te preocupa lo que habrá que hacer cuando llegue. Aunque no puedo demostrar que una autoridad suprema de la IA no vaya a aparecer algún día, sí puedo ofrecerte razones para que descartes con rigurosidad la perspectiva de ese escenario. Más en general, debes leer este libro si sientes curiosidad pero a la vez te encuentras confundido por el bombo generalizado que rodea a la IA en nuestra sociedad. Te explicaré los orígenes del mito de la IA, lo que sabemos y lo que ignoramos acerca de la perspectiva de alcanzar una IA de nivel humano, y el motivo por el que deberíamos apreciar mejor la única inteligencia verdadera que conocemos: la nuestra.

EN ESTE LIBRO

En la primera parte, «El mundo simplificado», explico que la cultura de la IA nos ha llevado a simplificar nuestras ideas sobre la gente a la vez que expandía nuestro conocimiento acerca de la tecnología. Esto comenzó con el fundador de la IA, Alan Turing, e incluye una serie de simplificaciones comprensibles pero desafortunadas que yo denomino «errores de inteligencia». Esos errores iniciales fueron magnificados hasta acabar conformando una ideología por parte de un amigo de Turing, el estadístico I. J. Good, quien introdujo la idea de «ultrainteligencia» como resultado predecible tras la consecución de una IA de nivel humano. Entre Turing y Good vemos cobrar forma al mito moderno de la IA. Su desarrollo nos ha conducido a una época de lo que yo llamo «tecnología kitsch», imitaciones baratas de ideas más profundas que anulan el compromiso inteligente y debilitan nuestra cultura. Lo kitsch nos indica lo que hemos de pensar y lo que hemos de sentir. Los proveedores del kitsch sacan rédito de él, mientras que los consumidores de ese kitsch experimentan una pérdida; acaban —acabamos— metidos en un mundo de frivolidad.

En la segunda parte, «El problema de la inferencia», argumento que no tenemos la menor idea sobre cómo programar o diseñar el único tipo de inferencia —de pensamiento, en otras palabras— que funcionará con una IA de nivel humano (o

cualquier otra cosa que se le acerque). El problema de la inferencia apunta al corazón del debate sobre la IA porque trata directamente con la inteligencia, la de la gente o la de las máquinas. Nuestro conocimiento acerca de los distintos tipos de inferencia se remonta a Aristóteles y a otros griegos de la Antigüedad, y se ha desarrollado en los ámbitos de la lógica y de las matemáticas. La inferencia ya se describe usando sistemas formales y simbólicos como los programas informáticos, así que explorándola se puede obtener una visión muy clara del proyecto con el que diseñar la inteligencia. Hay tres tipos de inferencia. La IA clásica exploró uno (las deducciones), la IA moderna explora otro (las inducciones). Y el tercer tipo (las abducciones) conduce a la inteligencia general y, sorpresa: nadie está trabajando en él —nadie en absoluto—.[1] Por último, puesto que todos los tipos de inferencia son distintos —con ello quiero decir que ninguno de esos tipos puede rebajarse hasta convertirse en otro—, sabemos que un fracaso a la hora de construir sistemas de IA que usen el tipo de inferencia en el que se afianza la inteligencia general conducirá al fracaso de los avances hacia la inteligencia artificial general, o IAG.

En la tercera parte, «El futuro del mito», argumento que, cuando se lo toma uno en serio, el mito tiene consecuencias muy negativas, ya que subvierte la ciencia. En especial, erosiona la cultura de la invención y la inteligencia humanas, que resultan necesarias en aquellos descubrimientos imprescindibles para comprender nuestro propio futuro. La ciencia de datos (la aplicación de la IA a los macrodatos) es, en el mejor de los casos, una prótesis del ingenio humano; en caso de usarla de manera correcta, nos ayudará a lidiar con el «diluvio de datos» contemporáneo. Cuando se la usa para reemplazar la inteligencia individual, tiende a estropear la inversión sin ofrecer ningún resultado. Explico, en especial, que el mito ha afectado negativamente la investigación en neurociencia, entre otros avances científicos recientes. Estamos pagando un precio demasiado elevado por este mito. Como no poseemos ninguna buena razón científica para creer que el mito pueda hacerse realidad, puesto que contamos con todos los motivos para rechazarlo a fin de alcanzar la prosperidad en el futuro, tenemos que repensar de manera radicalla conversación sobre la IA.

Primera parte

EL MUNDO SIMPLIFICADO

Capítulo 1

El error de la inteligencia

La historia de la inteligencia artificial comienza con las ideas de una persona que contó con una enorme inteligencia humana: el pionero de la informática Alan Turing.

En 1950, Turing publicó un artículo provocador, «Maquinaria computacional e inteligencia», sobre la posibilidad de crear máquinas inteligentes.[1] Fue un texto audaz, que llegó en un momento en el que los ordenadores eran novedosos pero insignificantes, según los parámetros de hoy en día. Aquellas piezas pesadas y lentas de hardware servían para acelerar cálculos científicos como el del análisis criptográfico. Tras una larga preparación, se les podían proporcionar fórmulas de física y unas condiciones iniciales, y obtener de forma automática el radio de una explosión nuclear. IBM no tardó en entender su potencial de cara a reemplazar a los seres humanos en sus operaciones comerciales, como la actualización de hojas de cálculo. Pero ver los ordenadores como criaturas «pensantes» requería de cierta imaginación.

La propuesta de Turing se basaba en un entretenimiento popular llamado «el juego de la imitación». En el juego original, un hombre y una mujer se ocultan a la vista y una tercera persona, el interrogador, les va haciendo preguntas alternativamente. A través de la lectura de sus respuestas tiene que determinar quién es el hombre y quién la mujer. La gracia está en que el hombre tiene que intentar engañar al interrogador, mientras que la mujer se esfuerza por ayudarle, lo cual conduce a que las respuestas de uno y otro lado resulten sospechosas. Turing reemplazó al hombre y a la mujer por un ordenador y una persona. Así nació lo que hoy conocemos como el «test de Turing»: un ordenador y una persona reciben las preguntas mecanografiadas de un juez humano, y si ese juez no logra identificar debidamente quién es el ordenador, el ordenador gana. Turing argumentó que, a partir de ese resultado, no dispondremos de ningún buen motivo para afirmar que la máquina carezca de inteligencia, sin importar que esta sea humana o no. Así, la cuestión de que la máquina disponga de inteligencia reemplazó la cuestión sobre si la máquina puede pensar de verdad.

El test de Turing, en realidad, es muy difícil: ningún ordenador lo ha superado. Por supuesto, en 1950 Turing desconocía este resultado a largo plazo; no obstante, al reemplazar las preguntas filosóficas problemáticas sobre la «consciencia» y el «pensamiento» con un test de resultados observables alentó la visión de la IA como una ciencia legítima con un objetivo bien definido. Mientras la IA cobraba forma durante los años cincuenta, muchos de sus pioneros y seguidores coincidieron con Turing: todo ordenador que pudiera mantener una conversación sostenida y convincente con una persona estaría, tal y como reconoceríamos la mayoría de nosotros, haciendo algo para lo que es necesario el pensamiento (sea eso lo que sea).

LA INTUICIÓN DE TURING / EL INGENIO COMO DISTINCIÓN

Turing se había labrado una reputación como matemático mucho antes de comenzar a escribir sobre IA. En 1936 publicó un artículo corto sobre el significado concreto de la palabra «computador», que en aquel momento se refería a la persona que seguía una serie de pasos para obtener un resultado definido (como la realización de un cálculo).[2] En aquel artículo reemplazó al computador humano por la idea de una máquina que realizara el mismo trabajo. El texto se adentraba en unas matemáticas de gran dificultad. Pero, mientras se refería a las máquinas, no hacía ninguna referencia al pensamiento humano ni a la mente. Las máquinas pueden operar de manera automática, afirmaba Turing, y los problemas que solucionan no requieren de ninguna ayuda «externa» o inteligencia. Esa inteligencia externa —el factor humano— es lo que los matemáticos a veces denominan «intuición».

El trabajo que Turing dedicó en 1936 a las máquinas computadoras ayudó a lanzar la ciencia informática como disciplina, y representó una contribución importante a la lógica matemática. Aun así, al parecer Turing pensó que aquella definición temprana pasaba por alto una cuestión esencial. De hecho, la misma idea de que la mente o las facultades humanas pudieran ayudar a solucionar problemas apareció dos años después en su tesis doctoral, un inteligente pero fallido intento de esquivar uno de los resultados obtenidos por Kurt Gödel, matemático de origen austriaco especializado en lógica (volveremos a él en un

rato). La tesis de Turing contiene este curioso pasaje sobre la intuición, que compara con otra capacidad mental a la que llama «ingenio»:

El razonamiento matemático puede ser considerado, de manera bastante esquemática, como un ejercicio de combinación entre dos facultades, a las que podríamos denominar intuición e ingenio. La actividad de la intuición consiste en realizar juicios espontáneos que no son el resultado de un hilo de razonamientos conscientes. Esos juicios son a menudo correctos, pero de ninguna manera lo son siempre (dejando de lado la cuestión sobre lo que se quiera decir con «correcto»). A menudo resulta posible encontrar otra manera de verificar la corrección de un juicio intuitivo. Por ejemplo, se puede juzgar que todos los números enteros positivos son factorizables en números primos; la argumentación matemática detallada conducirá a idéntico resultado. Esta también incluirá juicios intuitivos, pero serán menos susceptibles a la crítica que el juicio original sobre la factorización. No pretendo explicar esta idea de «intuición» de manera más explícita».

A continuación, Turing pasa a explicar el ingenio:

En matemáticas, el ejercicio del ingenio consiste en apoyar la intuición a través de una disposición adecuada de las proposiciones, y quizá de las figuras geométricas o de los dibujos. Lo que se pretende es que, cuando estos se encuentren dispuestos de manera verdaderamente correcta, la validez de los pasos intuitivos que sean necesarios no pueda ser motivo de una duda seria.[3]

Aunque su lenguaje se dirija a los especialistas, Turing señala lo evidente: por lo general, los matemáticos escogen sus problemas o «ven» un problema de interés en el que trabajar sirviéndose de una habilidad que cuando menos parece no poder dividirse en pasos, y que, por tanto, no se presta con claridad a la programación informática.

LA PERCEPCIÓN DE GÖDEL

También Gödel pensaba en la inteligencia mecánica. Igual que Turing, estaba obsesionado con la diferencia entre «ingenio» (mecánica) e «intuición» (mente). La distinción que él realizaba era en esencia la misma que la de Turing, aunque con un lenguaje diferente: demostración frente a verdad (o «teoría de la demostración» frente a «teoría de los modelos», en la jerga matemática). ¿Son, en definitiva, los de demostración y verdad el mismo concepto?, se preguntó Gödel. En caso afirmativo, las matemáticas e incluso la ciencia misma podrían entenderse de manera exclusivamente mecánica. Según esa visión, el pensamiento humano también sería mecánico. El concepto de AI, aunque el término no se hubiera acuñado aún, flotaba sobre la cuestión. ¿Se puede reducir la intuición de la mente, su capacidad para captar la verdad y el significado, a una máquina, a la computación?

Esta era la pregunta de Gödel. Al intentar contestarla, se encontró con un obstáculo que no tardaría en darle fama mundial. En 1931, Gödel publicó dos teoremas de lógica matemática conocidos como los teoremas de incompletitud. En ellos demostró las limitaciones inherentes a todos los sistemas matemáticos formales. Fue un golpe brillante. Gödel demostró de manera inconfundible que las matemáticas —toda la matemática, con ciertas suposiciones directas— no son, hablando en sentido estricto, ni mecánicas ni «formalizables». De manera más específica, Gödel demostró que en todo sistema formal (matemático o informático) han de existir proposiciones Verdaderas, con uve mayúscula, pero que no se pueden comprobar dentro del sistema mismo, sirviéndose de alguna de sus normas. La mente humana puede reconocer esa proposición Verdadera, pero el sistema en el que se ha formulado no la puede demostrar (cosa que sí es demostrable).

¿Cómo alcanzó Gödel esa conclusión? Los detalles son técnicos y complicados, pero la idea básica de Gödel es que podemos tratar un sistema matemático lo bastante complejo para realizar sumas igual que un sistema de significado, casi como si fuera una lengua natural del estilo del inglés o el francés —y lo mismo sirve para todos los sistemas de mayor complejidad—. Al tratarlo de esa manera, posibilitamos que el sistema hable sobre sí mismo. Y puede contarnos, por ejemplo, que presenta ciertas limitaciones. Esa fue la percepción de Gödel.

Los sistemas formales, como los que aparecen en las matemáticas, permiten la expresión precisa de verdades y falsedades. Por lo general, establecemos lo que es verdad utilizando las herramientas de la demostración —nos servimos de unas reglas para demostrar algo y así sabemos que es indudablemente cierto. Pero ¿hay proposiciones verdaderas que no se puedan demostrar? ¿Puede la mente saber cosas que se le escapen al sistema? En el sencillo caso de la aritmética, expresamos verdades escribiendo ecuaciones como «2 + 2 = 4». Las ecuaciones básicas son proposiciones verdaderas dentro del sistema aritmético, demostrables según las normas de la aritmética. Aquí se da una equivalencia entre lo demostrable y lo verdadero. Antes de Gödel, los matemáticos pensaban que la matemática entera presentaba esa propiedad. Eso implicaba que las máquinas podrían producir en serie todas las verdades de los diferentes sistemas matemáticos limitándose a aplicar las normas de manera correcta. Es una idea hermosa, pero no es cierta.

A Gödel se le ocurrió la extraña pero poderosa propiedad de la autorreferencia. Se puede formar una versión matemática de expresiones autorreferenciales como «Esta proposición no se puede demostrar dentro de este sistema» sin quebrantar las reglas de los sistemas matemáticos. Pero las denominadas «proposiciones autorreferenciales de Gödel» introducen contradicciones en la matemática: si son ciertas, son indemostrables. Si son falsas, puesto que afirman ser indemostrables, en realidad son ciertas. Lo verdadero significa falso, y lo falso, verdadero: es una contradicción.

Retomando el concepto de «intuición», nosotros, los seres humanos, podemos ver que, de hecho, la proposición de Gödel es verdadera, pero, por culpa del resultado de Gödel, sabemos también que las normas del sistema no pueden demostrarla —en efecto, el sistema se muestra ciego ante aquello que sus normas no alcanzan a cubrir—.[4] Lo que es verdad y lo que es demostrable se desmontan entre sí. Y es posible que pase lo mismo con la mente y la máquina. En cualquier caso, los sistemas puramente formales tienen sus límites. No pueden probar desde su propio lenguaje algo que es cierto. En otras palabras, nosotros podemos ver cosas que al ordenador se le escapan.[5]

El resultado de Gödel representó un duro golpe para la idea, popular en aquel momento, de que todas las matemáticas se podían convertir en operaciones basadas en normas que produjeran una verdad matemática tras otra. El Zeitgeist pertenecía al formalismo, no a la conversación sobre las mentes, los espíritus, las almas y demás. En el campo de las matemáticas, el movimiento formalista

señaló un giro más amplio de los intelectuales hacia el materialismo científico y, en particular, el positivismo lógico —un movimiento dedicado a erradicar la metafísica tradicional, como el platonismo, con esas formas abstractas que no se podían percibir con los sentidos, y las nociones tradicionales de la religión, como la existencia de Dios—. En efecto, el mundo estaba orientándose hacia la idea de las máquinas de precisión. Y nadie abrazó la causa formalista con tanto vigor como el matemático alemán David Hilbert.

EL DESAFÍO DE HILBERT

A principios del siglo XX (antes de Gödel), David Hilbert había lanzado un desafío al mundo matemático: demostrar que la totalidad de las matemáticas descansaba sobre un fundamento seguro. La ansiedad de Hilbert era comprensible. Si las normas puramente formales de la matemática no podían demostrar todas y cada una de sus verdades, al menos en teoría era posible que las matemáticas escondieran contradicciones y paparruchas. Que hubiera una contradicción oculta en algún lugar de las matemáticas lo arruinaba todo, porque a partir de una contradicción se puede demostrar cualquier cosa. Y, por tanto, el formalismo ya no servía para nada.

Hilbert expresó el sueño de cualquier formalista: demostrar al fin que las matemáticas eran un sistema cerrado y regido solo por normas. La verdad era tan solo una «demostración». Adquirimos conocimiento cuando nos limitamos a rastrear el «código» de una demostración y confirmamos que no se ha violado ninguna norma. El sueño más amplio de Hilbert, apenas disfrazado, apuntaba en realidad a una cosmovisión, a una imagen del universo en que este mismo fuera un mecanismo. La IA comenzó a cobrar forma como idea, una postura filosófica que también podía demostrarse. El formalismo trataba la inteligencia como si fuera un proceso reglado. Una máquina.

Hilbert lanzó su desafío durante el Segundo Congreso Internacional de Matemáticos, que se celebró en París en 1900. El mundo intelectual le dedicó su atención. El desafío constaba de tres partes principales: demostrar que las matemáticas eran una disciplina completa; demostrar que las matemáticas eran una disciplina consistente y demostrar que las matemáticas eran una disciplina

decidible.

Con la publicación de sus teoremas de incompletitud, en 1931, Gödel hirió de muerte las partes primera y segunda del desafío de Hilbert. La cuestión de la decidibilidad quedó sin respuesta. Un sistema es decidible cuando existe un procedimiento definido (una demostración o una secuencia de pasos deterministas y evidentes) para establecer si una proposición construida a partir de las normas de ese sistema es verdadera o falsa. La proposición 2 + 2 = 4 tiene que ser Verdadera, y la proposición 2 + 2 = 5 tiene que ser Falsa. Y sucede lo mismo con todas las proposiciones que se puedan realizar con validez utilizando los símbolos y las reglas del sistema. Puesto que se creía que la aritmética era la base de las matemáticas, demostrar que las matemáticas eran decidibles implicaba demostrar el resultado de la aritmética y sus extensiones. Eso equivaldría a decir que los matemáticos, al «jugar» su partida con reglas y símbolos (la idea formalista), participaban de hecho en un juego válido que nunca conduciría a la contradicción ni al absurdo.

Turing quedó fascinado por el resultado de Gödel, que demostraba no el poder de los sistemas formales, sino más bien sus limitaciones. Se puso a trabajar en la parte que quedaba del desafío de Hilbert y comenzó a pensar en serio si podía existir un proceso de decisión para los sistemas formales. En 1936, con un artículo titulado «Números computables», demostró que no era así. Turing se dio cuenta de que el uso de la autorreferencia por parte de Gödel también podía aplicarse a las preguntas sobre los procesos de decisión, o, en efecto, a los programas informáticos. En especial, se percató de que debían de existir números (reales) que ningún método definido pudiera «calcular» al escribir su expansión decimal, dígito a dígito. Importó un resultado del matemático del siglo XIX Georg Cantor, quien había demostrado que los números reales (aquellos con expansión decimal) eran más numerosos que los enteros, por más que tanto los números reales como los enteros fueran infinitos. Es posible que Turing se subiera sobre hombros de gigantes, pero, al final, su labor en «Números computables» demostró una imposibilidad. Fue un resultado restrictivo: no era posible ningún proceso de decisión universal. En otras palabras, las reglas —incluso en matemáticas— no bastan. Hilbert se había equivocado.[6]

LO QUE IMPLICÓ PARA LA IA

Lo importante de cara a la IA es lo siguiente: Turing refutó que las matemáticas fueran decidibles inventando una máquina, una máquina determinista, que no requería de ninguna intuición o inteligencia para resolver problemas. Hoy en día nos referimos a esa formulación abstracta de una máquina como la máquina de Turing. Ahora mismo estoy tecleando en una de ellas. Las máquinas de Turing son los ordenadores. Que el marco teórico de la informática se implementara como idea colateral, como un medio para obtener un fin diferente, es una de las grandes ironías de la historia intelectual. Mientras trabajaba para refutar que las matemáticas mismas fueran decidibles, Turing fue el primero en inventar algo preciso y mecánico: el ordenador.

En su tesis de 1938, Turing esperaba que los sistemas formales fueran ampliables incluyendo normas adicionales (y a continuación conjuntos de normas, y conjuntos de conjuntos de conjuntos de normas) que pudieran resolver el «problema de Gödel». Descubrió, en cambio, que aquel sistema nuevo y más potente tendría un problema de Gödel nuevo y más complejo. No había manera de sortear la incompletitud de Gödel. No obstante, enterrada bajo las complejidades del razonamiento de Turing sobre los sistemas formales, había una extraña sugerencia que resultaba relevante de cara a la posible existencia de la IA. ¿Y si la facultad de la intuición no se podía reducir a un algoritmo, a las normas de un sistema?

En su tesis de 1938, Turing intentaba encontrar una salida para el resultado restrictivo de Gödel, pero descubrió que era imposible. En su lugar cambió de marcha, se puso a explorar la manera en que, en sus propias palabras, podría «reducir en gran medida» el requisito de la intuición humana a la hora de realizar cálculos. Su tesis tomó en consideración el poder del ingenio al crear sistemas de normas cada vez más complicados (resultó que el ingenio podía volverse universal: hay máquinas capaces de tomar como referencia a otras máquinas y así dirigir todas las que se puedan construir. Esta percepción, técnicamente una máquina de Turing universal en vez de una simple, iba a convertirse en el ordenador digital). Pero, en su trabajo formal sobre la computación, Turing se había ido de la lengua (quizá de manera involuntaria). Al permitir que la intuición fuera diferente y externa respecto a las operaciones de un sistema puramente formal como es el ordenador, Turing estaba, de hecho, sugiriendo que podían existir diferencias entre los programas de ordenador dedicados a las matemáticas y los matemáticos.

Por tanto, fue curioso el giro que Turing realizó entre sus primeros trabajos de los años treinta y la especulación de amplio espectro acerca de la posible aparición de ordenadores inteligentes en «Maquinaria computacional e inteligencia», que se publicó una década larga después. Hacia 1950, el debate sobre la intuición había desaparecido de los textos de Turing sobre las implicaciones de Gödel. Su interés se trasladó, en efecto, a la posibilidad de que los mismos ordenadores se convirtieran en «máquinas intuitivas». En esencia, decidió que el resultado de Gödel no era aplicable al asunto de la IA: si los seres humanos somos ordenadores muy avanzados, el resultado de Gödel solo implica que hay algunas proposiciones que no podemos comprender o ver como verdaderas, tal y como sucede con otros ordenadores menos complejos. Esas proposiciones podrían ser complejas e interesantes a extremos fantásticos. O tal vez fueran banales pero abrumadoramente complejas. El resultado de Gödel dejaba abierta la cuestión de si la mente no era más que una máquina de gran complejidad, con unas limitaciones muy complejas.

En otras palabras, la intuición había pasado a formar parte de las ideas de Turing acerca de las máquinas y sus poderes. El resultado de Gödel no podía afirmar (según Turing, en cualquier caso) que la mente fuera una máquina o no. Por un lado, la incompletitud sostiene que algunas proposiciones pueden entenderse como verdaderas desde el uso de la intuición, pero que eso no se puede demostrar a partir de un ordenador que se sirva del ingenio. Por el otro, un ordenador más poderoso puede utilizar axiomas (o más bits de código relevante) y demostrar el resultado, mostrando así que la intuición no está lejos de la computación en lo que a este problema se refiere. La cosa se convierte en una carrera armamentística: un ingenio cada vez más poderoso que sustituya a la intuición en problemas cada vez más complejos. Nadie puede anticipar quién ganará la carrera, así que nadie puede argumentar nada —usando el resultado de la incompletitud— sobre las diferencias inherentes entre intuición (la mente) e ingenio (la máquina). Pero, tal y como Turing sin duda sabía, de ser eso cierto, también lo sería al menos la posibilidad de una inteligencia artificial.

Así, entre 1938 y 1950, Turing cambió de opinión acerca del ingenio y la intuición. En 1938, la intuición era el misterioso «poder de selección» que ayudaba a los matemáticos a decidir los sistemas con los que debían trabajar y los problemas que debían resover. La intuición no era algo que se encontrara en el ordenador. Era algo que tomaba decisiones acerca del ordenador. En 1938, Turing no creía que la intuición formara parte de sistema alguno, lo cual sugería no solo que la mente y la máquina eran fundamentalmente diferentes, sino que

una IA paralela al pensamiento humano resultaba casi imposible.

Sin embargo, para 1950 había cambiado de parecer. Con el test de Turing, desafió a los expertos e hizo una especie de defensa de la intuición en las máquinas; fue como si preguntara: «¿Por qué no?». Aquello supuso un cambio radical. Parecía que una nueva visión de la inteligencia comenzaba a cobrar forma.

¿Por qué ese cambio? Entre 1938 y 1950, a Turing le pasó algo ajeno al ámbito de las matemáticas estrictas y la lógica y los sistemas formales. Fue algo que le pasó, de hecho, a toda Gran Bretaña, y ciertamente a la mayor parte del mundo. Lo que pasó fue la segunda guerra mundial.

Capítulo 2

Turing en Bletchley

A Turing le fascinaba el juego del ajedrez —igual que a I. J. «Jack» Good, su colega matemático en tiempos de guerra. Cuando se enfrentaban (solía ganar Good), elaboraban procesos de decisión y reglas de oro para los movimientos ganadores. Jugar al ajedrez implica seguir las reglas del juego (ingenio), pero también parece requerir de cierta percepción (intuición) sobre las jugadas que pueden elegirse según las diferentes posiciones que se den sobre el tablero. Para ganar al ajedrez no basta con aplicar las reglas; en primer lugar, hay que saber qué reglas escoger.

Turing veía el ajedrez como una manera útil (y sin duda entretenida) de pensar sobre las máquinas y la posibilidad de conferirles intuición. Al otro lado del Atlántico, el fundador de la teoría de la información moderna, Claude Shannon, colega y amigo de Turing en Bell Labs, también pensaba en el ajedrez. Más adelante construyó uno de los primeros ordenadores que lo jugaron, una ampliación de la labor que había realizado anteriormente en un protoordenador llamado «el analizador diferencial», que podía convertir ciertos problemas de cálculo en procedimientos mecánicos.[1]

EL PRINCIPIO DE LA SIMPLIFICACIÓN DE LA INTELIGENCIA

El ajedrez fascinaba a Turing y a sus colegas en parte porque parecía que un ordenador podría programarse para jugar sin que la persona que lo programara necesitara saber todo por anticipado. Puesto que los dispositivos informáticos implementaban conectores lógicos como si-entonces, o e y, se podría ejecutar un programa (un conjunto de instrucciones) que generara resultados diferentes dependiendo de los escenarios con los que se encontrara mientras repasaba sus

instrucciones. Esa capacidad para cambiar de rumbo según lo que «viera» parecía, a juicio de Turing y sus colegas, simular un aspecto fundamental del pensamiento humano.[2]

Los jugadores de ajedrez —Turing, Good, Shannon y demás— tenían también en la cabeza otro problema matemático con una apuesta mucho más elevada. Trabajaban para sus gobiernos, ayudando a descifrar los códigos secretos que usaba Alemania para coordinar sus ataques contra los barcos comerciales y militares que cruzaban el canal de la Mancha y el océano Atlántico. Turing se comprometió con un esfuerzo desesperado por ayudar a derrotar a la Alemania nazi durante la segunda guerra mundial, y fueron sus ideas sobre computación las que contribuyeron a alterar el curso de la guerra.

BLETCHLEY PARK

Bletchley Park, sita de manera discreta en un pueblo pequeño y alejado del reguero de bombas que caían sobre Londres y la Gran Bretaña metropolitana, era un centro de investigación establecido para ayudar a descubrir la localización de los U-boote, los submarinos alemanes, que causaban estragos en las rutas marinas del canal de la Mancha. Los submarinos nazis representaban un problema capital para las fuerzas aliadas; habían hundido miles de embarcaciones y destruido enormes cantidades de suministros y equipamiento. Para mantener el esfuerzo de guerra, Gran Bretaña necesitaba importaciones de treinta millones de toneladas al año. En un momento dado, los U-boote llegaron a reducir esa cantidad en 200.000 toneladas al mes, siguiendo una estrategia de guerra reveladora y potencialmente catastrófica, para la que durante bastante tiempo no hubo réplica. En respuesta, el gobierno británico reunió a un grupo de criptoanalistas, jugadores de ajedrez y matemáticos talentosos para que investigaran la manera de descifrar las comunicaciones con los submarinos, conocidas como «cifrados». (Un «cifrado» es un mensaje oculto. Descifrar un mensaje consiste en convertirlo de nuevo en un texto legible.)[3]

Los códigos se generaban a través de un aparato con aspecto de máquina de escribir conocido como Enigma, que se comercializaba desde los años 1920 pero que los alemanes habían reforzado de manera importante para usarla en la

guerra. Las máquinas Enigma modificadas se utilizaron en todo tipo de comunicaciones estratégicas dentro del esfuerzo de guerra nazi. La Luftwaffe, por ejemplo, las usó en su gestión de la guerra aérea, y lo mismo hizo la Kriegsmarine en sus operaciones navales. En general, se consideraba que los mensajes encriptados con la máquina Enigma modificada eran indescifrables.

El papel que Turing desempeñó en Bletchley y su consiguiente ascenso a la categoría de héroe nacional después de la guerra es una historia que ya se ha contado muchas veces. (En 2014, una gran producción cinematográfica, The Imitation Game [Descifrando Enigma], dramatizó su trabajo en Bletchley, así como su rol consiguiente en el desarrollo de los ordenadores.) El mayor logro de Turing fue relativamente desaborido, según criterios matemáticos puros, porque explotó una vieja idea de la lógica deductiva. El método, al que él y otras personas se referían medio en broma como «turinguismo», se basó en eliminar amplios números de posibles soluciones para los códigos de Enigma encontrando combinaciones en las que hubiera contradicciones. Las combinaciones contradictorias son una imposibilidad; en un sistema lógico no puede darse «A» y «no A» a la vez, tal y como no podemos estar «en la tienda» y «en casa» al mismo tiempo. El turinguismo fue una idea ganadora, y se convirtió en un gran éxito en Bletchley. Logró lo que se había exigido a aquellos «jóvenes genios» recluidos en el laboratorio de ideas al acelerar el descifrado de los mensajes de Enigma. Otros científicos de Bletchley concibieron estrategias diferentes para descifrar los códigos.[4] Sus ideas se ponían a prueba con una máquina llamada Bombe —nombre burlón que provenía de una máquina polaca anterior, la Bomba, y que con toda probabilidad se inspiró en los ruiditos que esta realizaba al terminar cada uno de sus cálculos—. Pensemos en la Bombe como en un protoordenador, capaz de ejecutar diferentes programas.

Más o menos en 1943, el Eje perdió su ventaja bélica en beneficio de las fuerzas aliadas, y ello se debió en no poca medida al esfuerzo continuado de los descifradores de Bletchley. Aquel equipo obtuvo un éxito célebre, y sus miembros se convirtieron en héroes de guerra. Hicieron carrera. Bletchley, mientras tanto, también se reveló como un refugio para el pensamiento dedicado a la computación: la Bombe era una máquina que ejecutaba programas para resolver problemas que los seres humanos por sí mismos no podían solucionar.

¿MÁQUINAS INTUITIVAS? NO

En el caso de Turing, Bletchley desempeñó un papel capital de cara a que materializara sus ideas sobre la posibilidad de crear máquinas inteligentes. Igual que sus colegas Jack Good y Claude Shannon, Turing percibió el poder y la utilidad de sus «juegos mentales» como criptoanalistas durante la guerra: podían descifrar mensajes que de otro modo resultaban completamente opacos para los militares. Los nuevos métodos computacionales no solo resultaban interesantes para pensar en un juego de ajedrez automatizado, sino que podían, de manera bastante literal, hundir barcos de guerra.

Turing (una vez más) pensaba en una abstracción: mentes y máquinas, o la idea general de inteligencia. Pero había algo extraño en su visión de lo que aquello implicaba. En los años cuarenta, la inteligencia no era un rasgo que se atribuyera en general a los sistemas formales, como era el caso de la Bombe de Bletchley, una máquina descifradora puramente mecánica. Gödel había demostrado que, por norma, la verdad no podía reducirse a lo formal, en el sentido de que participara en un juego formal con un conjunto de reglas establecidas, pero recuerda que su demostración dejó abierta la cuestión de si una máquina específica podría incorporar la intuición de la que se sirve la mente para tomar decisiones sobre las reglas que se deben seguir, pese a que no pudiera existir ningún sistema supremo capaz de demostrarlo todo (tal y como el propio Gödel había revelado de manera tan definitiva en 1931).

Tras abandonar Bletchley, Turing dedicó cada vez más tiempo a la cuestión de si era posible construir una máquina que fuera lo bastante potente como para usar a la vez la intuición y el ingenio. El enorme número de combinaciones posibles que había que comprobar de cara a descifrar los códigos alemanes resultaba abrumador para la intuición humana. Pero unos sistemas que contaran con los programas adecuados podrían cumplir con esa tarea al simplificar aquellas vastas posibilidades matemáticas. Para Turing, eso sugería que la intuición podía cobrar cuerpo en las máquinas. En otras palabras, el éxito de Bletchley implicaba que quizá se pudiera construir una inteligencia artificial.

No obstante, para que esa línea de pensamiento cobrara sentido, había que decidirse por una idea concreta de «inteligencia». La inteligencia, tal y como la ejercen los seres humanos, debía ser reducible —analizable— según los términos de la capacidad de la máquina. En esencia, la inteligencia debía ser reducible a la forma de la resolución de problemas. Al fin y al cabo, en eso

consiste el juego del ajedrez y en eso consiste también descifrar un código.

Y ahí está: la mayor muestra de genio por parte de Turing, y también su mayor error, consistió en pensar que la inteligencia humana se limitaba a resolver problemas. Tanto si había explicitado las ideas sobre máquinas inteligentes de su «Maquinaria computacional e inteligencia» de 1950 durante los años de la guerra como si no, queda claro que la experiencia de Bletchley materializó su visión posterior sobre la IA, y queda claro que la IA, a su vez, siguió de cerca esa misma senda, aunque sin el autoanálisis que hubiera sido necesario.

Pero una mirada más atenta al éxito descodificador de Bletchley revela de manera inmediata una simplificación peligrosa en sus ideas filosóficas acerca del hombre y la máquina. Bletchley fue un sistema inteligente, resultado de la coordinación militar (incluyendo el espionaje y la inteligencia, así como la captura de naves enemigas); de la inteligencia social que se estableció entre los militares y los diversos científicos e ingenieros que había allí, y (como sucede con todo en esta vida) a veces fue también cuestión de pura suerte. Lo cierto es que, en cuanto realidad práctica, la máquina Enigma modificada por los alemanes era impenetrable por medios puramente mecánicos. Los alemanes eran conscientes de ello; se habían basado en argumentos matemáticos sobre las dificultades de la descodificación mecánica. En parte, el éxito de Bletchley se debió, irónicamente, a la tozuda confianza de los comandantes nazis en el carácter inexpugnable de los cifrados de Enigma —de modo que, en momentos cruciales, tras descubrir que ciertos mensajes habían sido descifrados, se negaron a modificar o reforzar las máquinas, echándoles la culpa a operaciones de espionaje encubiertas en vez de aceptar aquella derrota científica—. Pero la niebla de la guerra hace que se mezclen no solo diferentes tecnologías novedosas, sino nuevas formas de inteligencia humana y social. La guerra no es como el ajedrez.

Por ejemplo, al principio de la guerra, las fuerzas polacas recuperaron fragmentos importantes de comunicaciones de Enigma que más tarde revelaron pistas de valor incalculable para la labor de Bletchley. Los polacos habían usado esos fragmentos (junto con otros, procedentes de fuentes rusas) para desarrollar su propia Bombe, aunque más simple, en una fecha tan temprana como 1938. La versión muy mejorada de Turing a principios de 1940 —la Bombe que usaba su «turinguismo»— dependió de aquella primeriza labor polaca, facilitada por los hechos que tenían lugar en el campo de batalla. Turing también vio a su colega Gordon Welchman introducir cambios en su propio diseño, al que le añadió un

«tablero diagonal» para simplificar aún más la búsqueda de contradicciones,[5] como respuesta a las mejoras que los alemanes habían realizado en Enigma. Ahí había dos mentes humanas sirviéndose de la intuición, trabajando conjuntamente en sociedad.

Hubo otros acontecimientos en el teatro de la guerra que resultaron de importancia capital. El 8 de junio de 1940, un portaaviones británico se hundió delante de la costa noruega. Aquel ataque facilitó la localización de los U-Boote, si bien se cobró un precio elevado con los numerosos marineros que acabaron en el fondo del mar. Pocas semanas antes, a finales de abril de 1940, la patrullera alemana VP2623, un miembro de la flota especialmente devastador, había sido capturada con un tesoro de pruebas de Enigma en su interior. Las piezas que se necesitaban para resolver el puzle de Enigma estaban llegando a manos aliadas y se abrían camino hacia el grupo de Bletchley.

Por sí mismos, aquellos fragmentos resultaban por completo inadecuados para descifrar con rapidez el futuro de las comunicaciones alemanas; para los criptoanalistas de Bletchley no eran más que «conjeturas», según la definición de un biógrafo de Turing. Pero facilitaron un primer paso de radical importancia a la hora de dar con la manera de programar las máquinas Bombe. Turing y sus colegas lo denominaron «la ponderación de las pruebas», tomando prestado un término que acuñó el científico y lógico norteamericano C. S. Peirce (quien ocupará un papel destacado en la segunda parte de este libro).[6]

Los matemáticos interpretan el peso de la evidencia de maneras diferentes, pero, en el caso del éxito de Bletchley (y para asuntos más amplios relacionados con la IA), equivale a aplicar conjeturas informadas, o intuiciones, para dirigir el ingenio, o las máquinas. Un fragmento de texto descifrado procedente de un submarino capturado puede significar cualquier cosa, tal y como una bola blanca hallada cerca de una bolsa de bolas blancas puede significar cualquier cosa, pero en cada caso podemos realizar suposiciones inteligentes para comprender lo que ha sucedido. Pensamos que resulta muy probable que la bola blanca haya salido de esa bolsa, pese a no haber visto que la sacaran de ella. Se trata de una suposición. No se puede demostrar que ese tipo de suposiciones sean ciertas, pero, cuanto mejor funcione la intuición humana al establecer las condiciones iniciales para trazar los procesos mecánicos, mejores serán las posibilidades de que esos procesos acaben obteniendo los resultados deseados en vez de, pongamos por caso, prolongarse sin rumbo fijo, siguiendo direcciones erróneas o engañosas. El peso de la evidencia —suponer— hizo que las Bombe

funcionaran.

Los científicos de Bletchley no se limitaron a proveer a las Bombe de información, dejándolas luego para que realizaran la labor incansable e importante de eliminar millones de códigos o cifrados incorrectos. Desde luego, las Bombe fueron necesarias —eso es lo que Turing entendió con gran claridad, y lo que sin duda inundó su imaginación con la posibilidad de que aquellos «procesos mecánicos» pudieran reproducir o reemplazar a la inteligencia humana—, pero la realidad fue que el grupo de Bletchley se ocupó ante todo de hacer conjeturas. Al reconocer las pistas escondidas en el mosaico de instrucciones incompletas, cifrados y mensajes procedentes del campo de batalla, pasaron a generar hipótesis. En la ciencia, las conjeturas se definen así, como la formación de hipótesis (concepto que también utilizó Charles Sanders Peirce), y tienen una importancia fundamental para el progreso del saber humano. No es de extrañar, pues, que la obra de Bletchley equivaliera a un sistema de conjeturas acertadas. Su condición sine qua non no fue de tipo mecánico, sino que más bien podríamos describirla como una observación inicial inteligente. Las Bombe necesitaban que las apuntaran hacia algo, y que a continuación las impulsaran en ese sentido.

En sintonía con un tema que exploraremos en la segunda parte del libro, Peirce reconoció muy al principio, a finales del siglo XIX, que todas las observaciones que dan cuerpo a las ideas complejas y juicios de la inteligencia comienzan con una suposición, o lo que él llamó una abducción:

Al mirar por la ventana en esta hermosa mañana de primavera veo una azalea en plena floración. ¡No, no! No es eso lo que he visto, aunque sí se trate de la única manera en que puedo describirlo. Es una proposición, una frase, un dato; pero lo que percibo no es una proposición, una frase, un dato, sino apenas una imagen que yo hago inteligible en parte a través de la exposición de un hecho. Esa exposición es abstracta, pero lo que yo veo es concreto. Realizo una abducción cada vez que expreso cualquier cosa que haya visto en una frase. La verdad es que el entramado al completo de nuestro conocimiento es un fieltro opaco de hipótesis puras confirmadas y refinadas a través de la inducción. No se puede obtener el menor avance en el campo del conocimiento sin realizar una abducción a cada nuevo paso, o de otro modo nos quedaríamos mirando las cosas con expresión vacía.[7]

Turing y sus colegas de Bletchley comenzaron a ganar una guerra que había orbitado de los mandamases a los servicios de inteligencia gracias al uso, en efecto, de abducciones inteligentes en cada nuevo paso del camino. Hasta cierto punto, es evidente que Turing era consciente de ello (recordemos el debate sobre la intuición en la tesis que dedicó en 1938 a los números ordinales), pero no parece haber tenido un efecto apreciable en sus ideas posteriores sobre la naturaleza de la inteligencia y la posibilidad de crear máquinas inteligentes. Por brillante que se mostrara, formuló una simplificación de la inteligencia real. Se liberó del concepto que tanto le había subyugado con anterioridad: el de la intuición. El de las conjeturas.

SOBRE LA INTELIGENCIA SOCIAL

(UNA ACOTACIÓN IMPORTANTE)

La inteligencia social también quedó visiblemente fuera de la forma en que Turing resolvió el acertijo sobre la inteligencia. Esto es de la mayor importancia de cara a comprender el desarrollo futuro de la IA. Por ejemplo, a Turing le desagradaba considerar que el pensamiento o la inteligencia pudieran ser circunstancias sociales o situacionales.[8] Sin embargo, el éxito de Bletchley formó parte, en realidad, de un vasto sistema que se extendió mucho más allá de las cuatro paredes del lugar. Se había puesto en marcha un esfuerzo inmenso, que no tardaría en atraer a Estados Unidos y la labor de científicos como Shannon, en Bell Labs, así como de los que trabajaron en el célebre Instituto de Estudios Avanzados de Princeton —donde tenían puestos Einstein, Gödel y John von Neumann—. El sistema expandido de máquinas humanas resulta en realidad mucho más realista como modelo de la manera en que se solucionan los problemas del mundo real —entre ellos, el de una guerra mundial debe contarse sin duda como uno de los más complejos e importantes.

La falta de oído musical de la IA para la inteligencia social o situacional ya se había comentado antes, y en tiempos más recientes lo ha hecho el científico especializado en aprendizaje automático François Chollet, quien lo resume bien

en su crítica a la visión que tenía Turing sobre la inteligencia (y, de manera más amplia, a la del campo de la IA). Primero, la inteligencia es situacional, no existe nada parecido a una inteligencia general. Tu cerebro es una pieza dentro de un sistema más amplio que incluye tu cuerpo, tu entorno, a otros seres humanos y la cultura en su conjunto. Segundo, la inteligencia es contextual: lejos de existir en el vacío, cualquier inteligencia individual siempre se hallará definida a la vez que limitada por su entorno. (Y, en estos momentos, es el entorno, y no el cerebro, el que actúa como cuello de botella para la inteligencia.) Tercero, la inteligencia humana se encuentra en gran medida externalizada, contenida no en tu cerebro sino en tu civilización. Pensemos en los individuos como si fueran herramientas cuyos cerebros son módulos de un sistema cognitivo mucho más amplio que ellos mismos, un sistema que lleva mucho tiempo evolucionando.[9]

Según lo expresa Turing, la intuición se puede programar en una máquina, pero Chollet y críticos similares aseguran que esta no podrá alcanzar el nivel de la inteligencia humana. De hecho, la idea de programar la intuición ignora un aspecto fundamental de nuestros propios cerebros. Los seres humanos disponemos de inteligencia social. Disponemos de inteligencia emocional. Usamos nuestras mentes para algo más que para resolver problemas y acertijos, por complejos que sean (o, más bien, sobre todo cuando esos problemas son complejos).

La evidencia sugiere que Turing rechazó con firmeza esa visión de las personas, y en su lugar llegó a creer que la totalidad del pensamiento humano se podía entender, en efecto, desde el «desciframiento» de unos «códigos» —o resolución de acertijos— y la práctica de juegos como el ajedrez. Lo importante es que, en algún momento de los años cuarenta, después de trabajar en Bletchley y sin duda durante la época en la que escribió el artículo aparecido en 1950 donde prefiguraba la IA, el pensamiento de Turing se decantó por una visión simplificada de la inteligencia. Fue un error atroz, que además se ha ido transmitiendo de generación en generación de científicos de IA hasta llegar al día de hoy.

EL ERROR DE TURING CON LA INTELIGENCIA Y UNA IA DÉBIL

Esa visión de la inteligencia como algo que resuelve problemas ayuda a explicar la producción de aplicaciones invariablemente débiles a lo largo de la historia de la IA. Los juegos, por ejemplo, han sido una fuente constante de inspiración para el desarrollo de técnicas avanzadas de IA, pero estos no dejan de ser versiones simplificadas de la vida que recompensan visiones también simplificadas de la inteligencia. Un programa de ajedrez puede desempeñarse bien en ese juego, pero se le dará bastante mal conducir un coche. El sistema Watson de IBM juega al Jeopardy!, pero no al ajedrez ni al go, y se requiere un esfuerzo inmenso de programación o de «conversión» para que la plataforma Watson realice otras funciones de extracción de datos y procesamiento del lenguaje natural, como con sus recientes (y en gran medida fallidas) incursiones en el terreno de la salud.

Por consiguiente, tratar la inteligencia como algo que resuelve problemas conduce a que las aplicaciones de la IA sean débiles. Sin duda, Turing fue consciente de ello, y en su artículo de 1950 especuló con la posibilidad de que se pudiera hacer que las máquinas aprendieran y así superar las limitaciones que surgen como consecuencia natural del diseño de unos sistemas informáticos que solo sirven para solucionar problemas. Si las máquinas aprendieran a volverse genéricas, seríamos testigos de una transición fluida entre las aplicaciones específicas y unos seres dotados de pensamiento general. Llegaríamos a la IA.

No obstante, el conocimiento que tenemos hoy choca con violencia contra el enfoque de aprendizaje sugerido de manera temprana por Turing. Para alcanzar sus objetivos, los que en la actualidad denominamos «sistemas de aprendizaje automático» deben aprender algo específico. Los investigadores hablan de darle a la máquina un «sesgo» (sin las connotaciones negativas que le otorgamos en nuestra sociedad; no se pretende decir que la máquina sea cabezota o que cueste discutir con ella, ni que tenga motivaciones secretas, según el sentido habitual de la palabra). En el aprendizaje automático, el sesgo significa que el sistema está diseñado y puesto a punto para aprender algo. Pero, por supuesto, ese es precisamente el problema de producir aplicaciones débiles que resuelvan problemas. (Y es el motivo, por ejemplo, por el que los sistemas de aprendizaje profundo que usa Facebook para reconocer rostros humanos no han aprendido también a hacerte la declaración de la renta.)

Peor incluso, los investigadores se han dado cuenta de que darle a un sistema de aprendizaje automático un sesgo a la hora de aprender una tarea o aplicación concreta lleva a que tenga un rendimiento peor en otras tareas. Hay una correlación inversa entre el éxito de la máquina al aprender algo y que consiga

aprender otra cosa. Incluso tareas en apariencia similares presentan esa relación inversa en su desempeño. Un sistema informático que aprenda a jugar al go a nivel de campeonato no aprenderá además a jugar al ajedrez a ese mismo nivel. El sistema del go ha sido diseñado de manera específica, con un sesgo particular hacia el aprendizaje de las reglas del go. Su curva de aprendizaje, tal y como la llaman, sigue, por tanto, el tanteo conocido de ese juego en particular y, en relación con cualquier otro juego, pongamos el Jeopardy! o el ajedrez, se vuelve inútil —de hecho, no existe.

El sesgo en el aprendizaje automático se ha entendido por lo general como una fuente de errores de aprendizaje, un problema técnico. (También, al ajustarse al uso común del lenguaje, ha adoptado acepciones secundarias que ofrecen resultados involuntarios pero inaceptables por, pongamos, su carga racial o de género.) El sesgo en el aprendizaje automático puede introducir errores solo porque el sistema no «busca» ciertas soluciones en primer lugar. Pero, de hecho, el sesgo es necesario para el aprendizaje automático: forma parte de él.

Un célebre teorema conocido como «no free lunch» demuestra con exactitud lo que observamos de manera anecdótica al diseñar y construir un sistema de aprendizaje. El teorema sostiene que, al aplicarse sobre un problema arbitrario, cualquier sistema de aprendizaje libre de sesgo no obtendrá resultados mejores que los que proporciona el azar. Es una manera elegante de decir que los diseñadores de sistemas deben conferir a estos un sesgo de manera deliberada, para que aprendan su propósito. Tal y como señala el teorema, un sistema en verdad libre de sesgo no sirve para nada. Hay técnicas complicadas, como la del «preentrenamiento» con datos, que se sirven de métodos no supervisados que exponen los rasgos de los datos que hay que aprender. Todo ello forma parte integral de un aprendizaje automático exitoso. Lo que queda fuera del debate, no obstante, es que ajustar un sistema para que aprenda su propósito inculcándole el sesgo deseado implica que se vuelva restrictivo, en el sentido de que ya no podrá generalizarse a otros dominios. En parte, construir e implementar con éxito un sistema de aprendizaje automático lleva a que este no se encuentre libre de sesgo y no sea general, sino que se centre en un problema de aprendizaje particular. Visto así, la restricción se encuentra integrada hasta cierto punto en esos enfoques. El éxito y la restricción son las dos caras de una misma moneda.

Por sí solo, ese hecho ya arroja serias dudas sobre cualquier expectativa de progresión fluida entre la IA actual y la IA de nivel humano el día de mañana. La gente que asume que la ampliación de los métodos modernos de aprendizaje

automático, como el aprendizaje profundo, podrán formarse desde cero o aprender a ser tan inteligentes como los seres humanos, no comprende las limitaciones fundamentales ya conocidas. Admitir la necesidad de suministrar un sesgo a los sistemas de aprendizaje es equivalente a la observación por parte de Turing de que la mente humana debe suministrar percepciones matemáticas externas a los métodos formales, ya que el sesgo del aprendizaje automático está determinado, antes del aprendizaje, por sus diseñadores humanos.[10]

EL LEGADO DE TURING

Para resumir la cuestión, la visión de la inteligencia como algo que resuelve problemas genera de manera necesaria aplicaciones débiles y, por tanto, resulta inadecuada para los objetivos más amplios de la IA. Heredamos esa visión de la inteligencia de Alan Turing. (¿A cuento de qué, por ejemplo, usamos el término «inteligencia artificial» en vez de, quizá, el de «simulación de tareas humanas»?)[11] La genialidad de Turing consistió en deshacerse de los obstáculos y objeciones teóricas en el trayecto hacia la posibilidad de diseñar una máquina autónoma, pero con ello limitó el alcance y la definición de la inteligencia misma. No es de extrañar, pues, que la IA comenzara produciendo aplicaciones débiles de resolución de problemas y que haya seguido haciéndolo hasta el día de hoy.

Una vez más, a Turing le desagradaba la consideración del pensamiento o la inteligencia como algo social o situacional. Sin embargo, pese a su tendencia a entender la inteligencia humana como un proceso mecánico individual —lo que dio pie en los años cuarenta, con la aparición de los primeros ordenadores, a incontables menciones periodísticas al «cerebro mecánico»—, resulta evidente que la conversación acerca de la inteligencia implica siempre, y de manera necesaria, situarla en un contexto más amplio. La inteligencia general (no débil) del tipo que todos exhibimos a diario no se debe a ningún algoritmo que se esté ejecutando dentro de nuestras cabezas, sino que recurre a la totalidad del contexto cultural, histórico y social desde el que pensamos y actuamos en el mundo. La IA apenas habría avanzado si sus diseñadores hubieran abrazado un entendimiento de la inteligencia tan amplio y complejo —es cierto—. A la vez, a resultas de la simplificación realizada por Turing, hemos acabado usando

aplicaciones débiles y no tenemos ningún motivo para esperar otras más generales si antes no se produce una reconceptualización radical de lo que queremos decir al hablar de IA.

En su artículo de 1950, Turing anticipó algunas de esas dificultades al sugerir que se podía hacer que las máquinas aprendieran. No obstante, lo que sabemos en este momento (en oposición a la excitación reciente sobre el aprendizaje automático) es que el aprendizaje mismo es un tipo de resolución de problemas posibilitado solo por la introducción de un sesgo en el aprendiz; sesgo que de manera simultánea facilita el aprendizaje de una aplicación en particular mientras que limita el desempeño en otras aplicaciones. De hecho, los sistemas de aprendizaje también son sistemas restrictivos de resolución de problemas. Puesto que no existe ningún puente teórico entre esos sistemas tan restrictivos y la inteligencia general del tipo que exhibimos los seres humanos, la IA ha caído en una trampa. Una serie de errores tempranos en la comprensión de la inteligencia han conducido, en grado diverso, pero de manera inexorable, a un punto muerto teórico en el núcleo de la IA.

Consideremos de nuevo la distinción original de Turing entre intuición e ingenio. Para él, el problema de la IA consistía en que la intuición —aquello que viene suministrado por el diseñador del sistema— pudiera de hecho «introducirse» en la parte formal de ese sistema (el ingenio de la máquina) y crear así un sistema capaz de escapar a la maldición de lo restrictivo al servirse de la intuición para escoger sus propios problemas —para volverse cada vez más inteligente y aprender—. Hasta el momento, nadie ha conseguido hacer eso con ningún ordenador. Nadie tiene la menor idea, siquiera, de la manera en que se podría llegar a ello. Sabemos que los diseñadores usan la intuición desde fuera de los sistemas de IA para indicar a estos los problemas específicos que deben resolver (o que deben aprender a resolver). La cuestión de que los sistemas utilicen la intuición de manera autónoma apunta directamente al núcleo de lo que denominaré «el problema de la inferencia», sobre el que hablaremos en la segunda parte del libro.

En esa segunda parte también habrá muchas otras cosas que comentar sobre «la trampa de la restricción». Pero antes hay más terreno que cubrir en este primer apartado. A continuación, pasaremos al tema de la superinteligencia, otro error de la inteligencia y una extensión natural del primero.

Capítulo 3

El error de la superinteligencia

Jack Good, el colega descifrador de Turing, también quedó fascinado por la idea de una máquina dotada de inteligencia. Es evidente que Turing allanó su imaginación cuando ambos se encontraban en Bletchley, pero, más adelante, Good añadió un giro como de ciencia ficción a las ideas de Turing acerca de la posibilidad de una inteligencia de nivel humano en los ordenadores. La idea de Good era sencilla: si una máquina puede alcanzar el nivel de la inteligencia humana, también puede sobrepasar el mero pensamiento humano.

Good consideraba evidente que una especie de bucle de retroalimentación permitiría a las máquinas inteligentes examinarse a sí mismas y mejorar, lo que conduciría a la creación de máquinas aún más inteligentes y resultaría en una «explosión de inteligencia» desenfrenada. La explosión de inteligencia seguiría porque cada máquina haría una copia aún más inteligente de sí misma, con el resultado de una curva exponencial de inteligencia en las máquinas que no tardaría en sobrepasar a los grandes genios de la humanidad. Good lo denominó«ultrainteligencia»:

Defínase como ultrainteligente a aquella máquina que puede rebasar con amplitud las actividades intelectuales de cualquier hombre, por listo que este sea. Puesto que el diseño de las máquinas es una de esas actividades intelectuales, la máquina ultrainteligente podría diseñar máquinas aún mejores; es incuestionable que seguiría una «explosión de inteligencia» y que la inteligencia humana quedaría muy atrás. Así, la primera máquina ultrainteligente sería el último producto que la humanidad necesitaría inventar, contando con que la máquina se mostrara lo bastante dócil como para indicarnos la manera de mantenerla bajo control.[1]

El filósofo de Oxford Nick Bostrom iba a recuperar el tema de Good algunas décadas más tarde con un best seller titulado Superinteligencia: Caminos, peligros, estrategias, donde defendía la misma postura: que la consecución de la IA tendría como consecuencia el inicio de una inteligencia superior a la humana en un proceso cada vez más vertiginoso de automodificaciones. Con un lenguaje ominoso, Bostrom se hace eco del futurismo de Good acerca del advenimiento de las máquinas superinteligentes:

Ante la perspectiva de una explosión de inteligencia, nosotros, los seres humanos, somos como niños pequeños que juegan con una bomba. Esa es la disparidad entre la potencia de nuestro juguete y la inmadurez de nuestra conducta. La superinteligencia representa un desafío para el que no estamos preparados en este momento, y para el que no estaremos preparados durante mucho tiempo. No sabemos bien cuándo tendrá lugar la detonación, aunque si nos llevamos el artefacto a la oreja podemos oír un tictac amortiguado. En el caso de un niño que tuviera una bomba sin detonar entre las manos, lo más sensato sería que la dejara en el suelo con cuidado, que saliera rápidamente de la habitación y que llamara al adulto más cercano. Sin embargo, aquí nos encontramos no con un niño sino con muchos, y cada uno de ellos tiene acceso a un mecanismo de activación independiente. Las posibilidades de que todos demostremos el sentido común de deshacernos de algo tan peligroso parecen nimias. Siempre habrá algún idiota dispuesto a apretar el botón de ignición solo para ver qué pasa.[2]

Para Bostrom, la superinteligencia no es objeto de especulación ni una nebulosa, sino que se parece más a la llegada de las armas nucleares: un hecho consumado que tendrá consecuencias profundas y quizá nefastas para la raza humana. El mensaje está claro: no discutas si la superinteligencia va a llegar. Prepárate para su llegada.

¿Qué hemos de decir ante eso? El argumento de Good y Bostrom —la posibilidad de una máquina superinteligente— parece verosímil a primera vista. Pero, de manera poco sorprendente, nunca se especifica el mecanismo por el que la inteligencia de partida llevará a la superinteligencia. Good y Bostrom parecen tomarse la posibilidad de la superinteligencia como algo tan creíble y evidente

que esta no requiere de mayores explicaciones. Pero sí que las necesita; tenemos que comprender el «cómo».

Si imaginamos una simple mejora como la de un hardware superior, la propuesta resulta demasiado trivial y ridícula como para que le dediquemos más tiempo. Es probable que ni siquiera un creyente incondicional en la inexorabilidad del progreso como Ray Kurzweil vaya a rebajar la inteligencia hasta ese punto —no pensamos que, al añadirle memoria RAM a un MacBook, estemos volviéndolo más inteligente (de verdad)—. El aparato irá más rápido, y podrá cargar aplicaciones de mayor tamaño y tal. Pero, si por inteligencia entendemos algo interesante, esta tiene que hacer algo más complejo que cargar aplicaciones a gran velocidad. Esa parte más difícil de la inteligencia se queda sin comentar.

O supongamos que tomamos prestado el lenguaje del mundo de la biología (como la IA hace tan a menudo) y, a continuación, declaramos con seguridad que la capacidad computacional no involuciona, sino que evoluciona. Una mirada más profunda nos hará ver que ese argumento se ve afectado una vez más por una visión inadecuada e ingenua de la inteligencia. El problema —una omisión flagrante— es que no existe ninguna evidencia en el mundo biológico de que algún ser inteligente haya diseñado alguna vez una versión más inteligente de sí mismo. Los seres humanos somos inteligentes, pero a lo largo de la historia de la humanidad no hemos construido ninguna versión más inteligente de nosotros mismos.

Hay un requisito previo a la construcción de un cerebro más listo, y consiste en comprender el carácter cognitivo del que ya tenemos, en el sentido de que podemos imaginar escenarios, albergar pensamientos y sus conexiones, encontrar soluciones y descubrir nuevos problemas. Nos pasan cosas, razonamos a través de nuestras observaciones y de lo que ya sabemos, las respuestas brotan en nuestra cabeza. Todo ese zumbido de magia biológica sigue siendo opaco, la mayor parte de su «procesamiento» continúa pendiente de cartografiar. Y, sin embargo, llevamos milenios considerando e investigando nuestros procesos mentales y nuestras funciones cerebrales.

Siendo evidente que con nosotros no sucede así, ¿por qué debería una máquina mayormente inteligente desarrollar de golpe una percepción sobre sus propias capacidades cognitivas globales? Y, aunque lograra hacerlo, ¿cómo podría la máquina usar ese conocimiento para volverse más lista?

No es una cuestión de autosuperación. Podemos, por ejemplo, volvernos más inteligentes leyendo libros o yendo a la escuela; al formarnos posibilitamos un mayor desarrollo intelectual, etcétera. Todo eso no es motivo de controversia. Y nada de ello viene al caso. Un problema capital de las conjeturas acerca del aumento de inteligencia en los círculos de la IA es el de su carácter circular: hace falta una inteligencia (en apariencia general) para hacer crecer la inteligencia general. Si lo observamos con detenimiento veremos que no hay ninguna progresión lineal, solo misterio.

VON NEUMANN Y LAS MÁQUINAS AUTORREPLICANTES

Good introdujo a mediados de los años sesenta la idea de una IA evolutiva que condujera a la ultrainteligencia. Pero, casi dos décadas antes, John von Neumann ya había considerado esa idea y la había descartado. En una conferencia de 1948 en el Instituto de Estudios Avanzados de Princeton, Von Neumann explicó que, mientras que la reproducción humana a menudo mejora los «diseños» previos, resulta evidente que las máquinas con el cometido de diseñar máquinas nuevas y mejores se enfrentan a un escollo fundamental, ya que cualquier diseño para una máquina nueva tiene que aparecer especificado en la máquina madre. La máquina madre tendría que ser por necesidad más compleja que su creación, no menos: «La organización sintetizadora ha de ser por necesidad más compleja, de un orden superior, que la organización sintetizada», dijo.[3]

En otras palabras, Von Neumann señaló una diferencia fundamental entre la vida orgánica tal y como la conocemos y las máquinas que construimos. La predicción de la ultrainteligencia que había hecho Jack Good era un poco de ciencia ficción.

Von Neumann propuso que la máquina autorreplicante debía tener, como mínimo, ocho partes, incluyendo un órgano que reciba y transmita «estímulos», un órgano «de fusión» que conecte las partes entre sí, un órgano «de corte» que interrumpa esas conexiones y un «músculo» para poder desplazarse. A continuación, bosquejó mecanismos que pudieran generar mejoras cognitivas de manera verosímil y que incluyeran un elemento aleatorio, similar a las mutaciones biológicas, para que permitiera las modificaciones necesarias. Pero

Von Neumann pensó que, en vez de hacer avanzar el pensamiento de la máquina, cabía presumir que esas mutaciones azarosas harían «involucionar» las funciones y capacidades que se buscaba obtener. El resultado más probable sería una avería, el equivalente de una alteración letal: «De modo que, mientras que este sistema resulta extremadamente primitivo, en él se incluye el rasgo de una mutación hereditaria, llegándose al extremo de que una mutación azarosa será probablemente letal, pero podría ser no letal y hereditaria».

Para que las máquinas obtuvieran algo mejor de sus diseños, en esencia una mayor inteligencia, necesitarían que se añadiera un elemento creativo a sus órganos de estímulos y de fusión. A diferencia de lo que sucede con la evolución biológica, la idea no tendría que esperar millones de años, sino que requeriría que los sistemas madre contuvieran el chispazo prometeico necesario, lo que conduciría de manera más o menos directa a una mejora de los diseños. Aquello era ficción, pensaba Von Neumann. Tal y como comentó ante sus colegas de Princeton, no había ciencia o teorías de ingeniería que pudieran encontrarle un sentido. Von Neumann, que no era ningún ludita, había hecho explotar la «explosión de inteligencia».

Un defecto evidente en las predicciones sobre la explosión de inteligencia que conduzca a la superinteligencia es que ya disponemos de una inteligencia de nivel humano: somos humanos. Siguiendo la lógica de Good, deberíamos ser capaces de diseñar algo por encima de lo humano. Esto no es más que una reafirmación de los objetivos del ámbito de la IA, así que nos hemos metido en un círculo vicioso. Las personas que se dedican a la investigación en IA ya saben que el diseño de artefactos más inteligentes es un misterio, tal y como explicó Von Neumann. Trasladar ese misterio desde nuestra propia inteligencia a la inteligencia imaginaria de una máquina no nos ayuda en nada. Para desentrañar este tema un poco más, pensemos en una investigadora genial de IA a la que llamaremos Alice.

EXPLOSIONES DE INTELIGENCIA, LA IDEA MISMA

Supongamos que Alice es una científica de IA que tiene un vecino tostón llamado Bob. Bob dispone de sentido común, puede leer el periódico y mantener

una conversación normal (aunque quizá sea aburrida), así que se encuentra a una distancia sideral por encima de los mejores sistemas de IA surgidos del DeepMind de Google.

Alice trabaja para una empresa emergente (que no tardará en ser adquirida por Google) y quiere construir una IA que sea tan lista como Bob. Ha bosquejado dos sistemas, siguiendo el espíritu de los célebres Sistema 1 y Sistema 2 de Daniel Kahneman.[4] Se trata de concesiones a la intuición o metáforas que proporcionan un borrador para los tipos de problemas que habrá que resolver a fin de llegar a la inteligencia artificial general. En el contexto de Alice, los llamaremos Sistema X, de aptitud en tareas bien definidas como la participación en juegos (el ajedrez o el go), y Sistema Y, de inteligencia general. Este último incluye las aptitudes lectoras y conversacionales de Bob, pero también el área más turbia de nuevas ideas y percepciones.

A Bob se le da fatal el ajedrez y, de hecho, su sistema X resulta patético en comparación no solo con el de un sistema como AlphaGo, sino respecto a muchos otros seres humanos. Su memoria a corto plazo es peor que la de la mayoría de la gente; obtiene resultados pobres en las pruebas de inteligencia y le cuesta bastante resolver los crucigramas. En lo que respecta a su sistema Y, su inteligencia general muestra una llamativa falta de interés o de capacidad para el pensamiento novedoso o perspicaz. Bob no es de esos vecinos que reciben demasiadas invitaciones para cenar.

La estrategia de Alice consiste en comenzar diseñando una Máquina de Bob que esté a la altura de su inteligencia. Según su razonamiento, si logra crear una Máquina de Bob, esta podrá diseñar una versión más inteligente de sí misma, lo cual acabará conduciendo a una explosión de inteligencia. Bien, una vez más no olvidemos que diseñar una Máquina de Bob no es tarea sencilla, porque Bob cuenta con un Sistema Y —lo cual significa que ha solucionado el problema del razonamiento y del sentido común, y dispone de capacidades cognitivas generales—. Por ejemplo, podría superar el test de Turing. Y puede leer cuentos infantiles y la sección de deportes y resumirlos. Bob, por tanto, podría machacar a los mejores sistemas de comprensión del lenguaje natural de Google, como la herramienta de búsqueda semántica Talk to Books de Ray Kurzweil. Ese es el motivo por el que Alice está tan emocionada con su proyecto de Máquina de Bob: sería un gran avance para la IA.

La pregunta es: ¿cómo llegar hasta ese punto? El primer enfoque de Alice

consiste en maximizar las capacidades del Sistema X de la Máquina de Bob. Le proporciona una memoria de ordenador y acceso a la red a través de Google. Por desgracia, esta versión de la Máquina de Bob no tarda en demostrar la afirmación de Stuart Russell según la cual los superordenadores sin inteligencia real simplemente tardan menos en obtener las respuestas equivocadas.[5] La Máquina de Bob recuerda las cosas erróneamente y fracasa a la hora de hacer las preguntas adecuadas. Pese a todas las mejoras que va introduciendo en su Sistema X, Alice solo consigue que la máquina se muestre más competente en recordar y expectorar teorías descabelladas y en realizar declaraciones acerca del mundo con más datos, todos ellos mal utilizados y peor entendidos desde la perspectiva de un Sistema Y. Sí, la Máquina de Bob juega de fábula al ajedrez, pero sus aptitudes para el ajedrez hacen que Alice la considere menos interesante, ya que se da cuenta de que la máquina que ha creado no tiene ninguna posibilidad de diseñar una versión «más inteligente» de sí misma.

En un momento de iluminación, Alice se da cuenta de que Bob no podría diseñar una versión más lista de sí mismo. Por tanto, ¿cómo podría hacerlo la Máquina de Bob? El problema, piensa, es que la optimización del Sistema X no suministra los recursos necesarios al Sistema Y. La Máquina de Bob (igual que el propio Bob) debería cuantificar su inteligencia, evaluar sus limitaciones y la extensión de estas, y a continuación rediseñarse de forma activa a fin de volverse más lista allí donde las cosas cuentan y tienen importancia. ¡Pero sucede que en ese punto es precisamente donde la Máquina de Bob (igual que Bob) se muestra poco inteligente! La Máquina de Bob no puede hacer eso porque su Sistema Y carece de las capacidades de percepción, descubrimiento e innovación. Alice tiene que volver a la casilla de salida.

Entonces, Alice decide que la Máquina de Bob es demasiado estúpida para formar parte del proceso de arranque de una superinteligencia. (En un momento de pánico cerval, se le ocurre que esa lógica pone en peligro la empresa entera de llegar a una superinteligencia, pero se las arregla para reprimir esa ansiedad con rapidez.) Alice decide, en deferencia al fundador de la IA y al entusiasta departamento de mercadotecnia de su compañía, Ultra++, que en su lugar va a concentrarse en el diseño de una máquina tan inteligente como Alan Turing, a la que llamará Máquina de Turing.

Bien, asumiendo que Turing fuera más listo que Alice (aunque ¿quién podría asegurarlo?), ella no puede diseñar directamente esa Máquina de Turing; de hecho ya se había estrellado contra una pared de ladrillos al intentar resolver el

diseño de la Máquina de Bob. Decide comenzar con una máquina que sea tan lista como Hugh Alexander —colega de Turing en Bletchley Park y, en una ocasión, campeón de ajedrez en Cambridge—. Hugh Alexander era listo, realmente listo. Jugaba al ajedrez a nivel profesional y, aunque no logró demostrar el mismo nivel de percepción que Turing a la hora de descifrar el código Enigma, sí realizó contribuciones valiosas a ese empeño y se ganó el respeto de los demás descifradores de Bletchley —que no es poco—. La Máquina de Hugh deberá ser lo bastante lista para averiguar la manera de cablear una Máquina de Turing, ¡y una máquina del nivel de Turing sin duda será lo bastante lista como para volverse más lista aún!

Solo con descargarse en el móvil un código ajedrecístico de StockFish, Alice ya logra mejorar la aptitud de la Máquina de Hugh para el ajedrez (y eso que Hugh fue un campeón en esa disciplina). De manera similar, gracias a una calculadora le otorga a la Máquina de Hugh una capacidad aritmética perfecta, y le añade una memoria de superordenador, así como acceso a toda la información que Google pueda recuperar. El Sistema X es inmejorable y la Máquina de Hugh puede hacer cosas que Hugh Alexander, con toda su inteligencia aparente, nunca pudo llevar a cabo: jugar un ajedrez sobrehumano, realizar sumas imposibles y destacar en muchas otras cosas propias del Sistema X. El problema es que eso también podía hacerlo la Máquina de Bob. De hecho, Alice se da cuenta de que la Máquina de Bob y la Máquina de Hugh son con toda probabilidad equivalentes. Incluso se ve obligada a admitir (tras varios vasos de vino tinto) que abandonar la Máquina de Bob fue innecesario y contraproducente.

Después de otro vaso de vino y un cigarrillo, Alice apaga el móvil para silenciar los molestos mensajes de texto que le están mandando sus colegas de Ultra++ para preguntar por su descubrimiento inminente. La verdad, reflexiona, es que Bob no es el único que no puede diseñar una versión más lista de sí mismo; a Alice le pasa igual. En un momento de lucidez descubre que, cuanto más nos alejemos del Sistema X en dirección al Sistema Y, hacia la percepción y la innovación, más opaco se volverá el diseño. Turing, por ejemplo, podía juzgar su inteligencia en el ajedrez —perdía tanto contra Hugh Alexander como contra Jack Good—, pero no podía evaluar sus propias capacidades de Sistema Y. En un sentido muy real, su intelecto era una caja negra, y bajo ningún concepto podía medir su propia aptitud para el pensamiento original (signifique eso lo que signifique), no solo porque este carece de un plano detallado, por así decirlo, sino porque se asienta en el tiempo, en el plazo de una vida, y aún podría generar ideas nuevas e impredecibles. La inteligencia del Sistema Y de Turing no solo

resulta impredecible; en otras palabras, es inexplicable —quizás no ante alguien más listo que Turing (de nuevo: signifique eso lo que signifique), pero sin duda ante el propio Turing—. Y lo mismo sucede con el lerdo de Bob. ¿Cómo podría Alice, pues, desencadenar una explosión de inteligencia?

Efectivamente, la idea misma de la explosión de inteligencia lleva incorporada una premisa falsa, fácil de exponer a alguien tan ambicioso y perspicaz como Alice, una vez que decide tomársela en serio. Según esa hipótesis, una Máquina de Bob será tan inteligente como Bob. Bien, he aquí una idea: ve a pedirle a Bob que diseñe una versión ligeramente más lista de sí mismo y descubrirás que eso no se encuentra a su alcance. La cualidad esencial de la mente que hace que la IA resulte tan excitante imposibilita a su vez el supuesto lineal de una explosión de inteligencia. «Una vez que lleguemos al nivel de la inteligencia humana, el sistema puede diseñar una versión de sí mismo más inteligente que los seres humanos», se dice con esperanza. Pero es que ya disponemos de una inteligencia de «nivel humano»—somos humanos—. ¿Podemos nosotros hacer algo así? ¿De qué están hablando en realidad los promotores de la explosión de inteligencia?

Es otra manera de decir que los poderes de la mente humana exceden nuestra capacidad para mecanizarla, en el sentido de que se necesita «ampliarla» para pasar del AlphaGo a la Máquina de Bob y a la Máquina de Turing y más allá. La idea misma de la explosión de inteligencia no representa una candidata particularmente buena como Sistema Y para que la IA progrese camino de una inteligencia general.

LOS EVOLUCIONISTAS TECNOLÓGICOS

Muchos entusiastas de la IA se aferran a la tesis de la inevitabilidad (las máquinas superinteligentes van a llegar, hagamos lo que hagamos) porque toca temas evolutivos y, así, de manera muy conveniente, absuelve a los científicos de la responsabilidad individual de realizar hallazgos o desarrollar ideas revolucionarias. La inteligencia artificial se limitará a evolucionar, igual que nosotros. Podemos denominar a esas voces futurólogas y creyentes en la IA dentro de este terreno como «evolucionistas tecnológicos», o ET.

La visión ET es popular entre tecnólogos new age como Kevin Kelly, cofundador de Wired, quien sostiene en su libro del año 2010 What Technology Wants [«Lo que la tecnología quiere»] que la IA no llegará por obra de un «científico loco», sino simplemente como un proceso evolutivo del planeta, de forma bastante parecida a la evolución natural.[6] Según su punto de vista, el mundo se está «inteligentizando» (el neologismo es suyo) y cada vez van emergiendo formas de tecnología más y más complejas e inteligentes, y sin un diseño humano explícito.[7] Es posible que esos pensadores también conciban la World Wide Web, la Red de Extensión Mundial, como un cerebro gigante y en crecimiento. Los seres humanos, según esta perspectiva, devienen un eslabón en la cadena cósmica-histórica que se extiende hacia el futuro y la verdadera IA, momento en que nos quedaremos atrás o seremos asimilados.

La vida orgánica evoluciona con extrema lentitud, pero los ET perciben un progreso tecnológico cada vez más acelerado. Tal y como dice una famosa aseveración de Kurzweil, la tecnología sigue una curva de aceleración y se está volviendo cada vez más complicada, según una ley que él considera discernible a lo largo de la historia: la ley de rendimientos acelerados. Así, la inteligencia de nivel humano y, a continuación, la superinteligencia aparecerán en el planeta en un plazo de tiempo tremendamente corto si lo comparamos con la evolución orgánica. Faltan pocas décadas para que tengamos que enfrentarnos a ellas, quizá solo años.

Esta es una historia de la humanidad pulcra y simple. Nos encontramos en medio de una transición hacia algo diferente, algo que será mejor y más inteligente.

Date cuenta de que esa historia no es comprobable; tan solo tenemos que esperar y ver qué pasa. Si el año que se ha predicho para la llegada de la verdadera IA resulta ser también falso, pues se pronostica otro para dentro de algunas décadas. En ese sentido, la IA es infalsificable y por tanto —según las reglas aceptadas del método científico— carece de rigor, no es una idea científica.

Fijaos en que no estoy diciendo que la IA verdadera sea imposible. Como les gusta señalar a Stuart Russell y a otros investigadores de IA, algunos científicos del siglo XX, como Ernest Rutherford, pensaron que era imposible construir una bomba atómica, pero Leó Szilárd descubrió la manera en que operan las reacciones nucleares en cadena —y lo hizo apenas veinticuatro horas después de que Rutherford diera la idea por muerta—.[8] Es un buen recordatorio de que no se debe apostar contra la ciencia. Pero piensa que la reacción nuclear en cadena se

desarrolló a partir de unas teorías científicas comprobables. Las teorías acerca de la evolución tecnológica de un poder mental no lo son.

Las declaraciones de Good y Bostrom, presentadas como una inevitabilidad científica, son más bien una concesión a la fantasía: ¡imagínate que esto fuera posible! Y no cabe duda de que sería genial. Y quizá peligroso. Pero imaginar escenarios hipotéticos nos aleja mucho de una discusión seria sobre lo que nos espera.

Para comenzar, una capacidad de superinteligencia general debería estar conectada al resto del mundo de manera que pudiera observar y «hacer conjeturas» de manera más productiva que nosotros. Y, si la inteligencia también es social y situacional, tal y como parece que debe de ser, se requerirá una inmensa cantidad de conocimiento contextual para diseñar algo más inteligente. El problema de Good no es mecánico y restrictivo, sino que más bien atrae hacia su órbita la totalidad de la cultura y la sociedad. ¿Dónde está el plano más simple y remotamente plausible para ello?

En otras palabras, la propuesta de Good se basa, una vez más, en una visión de la inteligencia simplista e inadecuada. Presupone el error original de la inteligencia y le añade otro juego de manos reduccionista: que una inteligencia mecánica individual puede diseñar y construir otra inteligencia mecánica individual superior. Que una máquina pueda situarse en tamaño punto de creación arquimédica parece improbable, por decirlo con suavidad. En realidad, la idea de la superinteligencia es una multiplicación de errores, y representa la esencia del punto al que ha llegado la fantasía en relación con el advenimiento de la IA.

Para ahondar en todo esto tenemos que seguir abriéndonos paso hacia el interior de esa fantasía. Se conoce como la «singularidad», y nos vamos a centrar en ella a continuación.

Capítulo 4

La singularidad, ayer y hoy

En los años cincuenta, el matemático Stanisław Ulam recordaba una vieja conversación que mantuvo con John Von Neumann en la que este comentó la posibilidad de que la humanidad viviera un punto de inflexión tecnológico: «El proceso en aceleración constante de la tecnología ... parece sugerir que nos acercamos a una singularidad esencial en la historia de nuestra raza, tras la cual los asuntos humanos tal y como los conocemos no podrán seguir igual».[1]

Lo más probable es que Von Neumann efectuara ese comentario en un momento en que los ordenadores digitales estaban llegando a la escena tecnológica. Pero los ordenadores digitales fueron la última innovación dentro de una larga y en apariencia ininterrumpida secuencia tecnológica.[2] En los años cuarenta ya había quedado claro que las revoluciones científica e industrial de los trescientos años precedentes habían puesto en movimiento unas fuerzas de un enorme poder simbiótico: los frutos de la nueva ciencia sembraron el desarrollo de la nueva tecnología, que a su vez posibilitó mayores descubrimientos científicos. Por ejemplo, la ciencia nos dio el telescopio, que a su vez permitió grandes avances en astronomía.

Los cambios sociales —rápidos, a veces caóticos y al parecer irreversibles— estuvieron ligados de manera inextricable a los cambios en la ciencia y la tecnología. La población de las ciudades se disparó (con dosis considerables de miseria e injusticia) y, aparentemente de la noche a la mañana, emergieron formas por completo nuevas de organización social y económica. Las máquinas de vapor revolucionaron el transporte, igual que los motores de combustión interna más adelante. Trenes, tranvías y barcos de vapor abrieron nuevas rutas comerciales y la migración a las ciudades generó nuevas fuerzas de trabajo. Con la invención de la bombilla eléctrica por parte de Thomas Edison, la gente pudo trabajar de noche; en las zonas rurales, quienes padecían insomnio podían ahora leer El Capital o El origen de las especies cuando el sol ya los hubiera abandonado. La productividad subió como la espuma. La riqueza y la salud crecieron. Lo mismo pasó con la sangre y la violencia. Una secuencia de

acontecimientos geopolíticos condujo a la «Gran Guerra», la primera guerra mundial, que introdujo el uso de armas químicas a escala masiva. Y, un par de décadas más tarde, en el mundo de Von Neumann, la amenaza existencial definitiva —la bomba nuclear— se hizo realidad.

La bomba marcó un punto de inflexión histórico, dejó a las claras las posibilidades distópicas inherentes a la innovación tecnológica desenfrenada. Shannon y Turing usaban ordenadores para jugar al ajedrez; científicos como Von Neumann usaron ordenadores para desarrollar armas con las que vaporizar ciudades japonesas. Los ordenadores electrónicos eran grandes y lentos, pero seguían siendo considerablemente más rápidos que los ordenadores humanos para realizar tareas como el cálculo de progresiones numéricas, que Von Neumann y otros usaban para determinar el radio probable de una explosión nuclear según las diferentes cantidades de material fisible.

En este miasma de miedos y posibilidades, Von Neumann fue a plantear la cuestión de una «singularidad». Famoso polímata y científico brillante, Von Neumann disfrutaba de un respeto casi universal entre sus colegas, incluyendo a Alan Turing, y no es de extrañar que sus sugerencias impactaran a Ulam, quien habría de recordarlas décadas más tarde.

No cabe duda de que el matemático Ulam entendió la metáfora de Von Neumann. El de singularidad es un término matemático que indica un punto que se vuelve indefinido —un valor que, pongamos, estalla hacia el infinito—. Von Neumann le preguntó a Ulam si el progreso tecnológico en efecto se acercaría a ese «infinito», sobre el que no se podrían aplicar métodos ni ideas, estrategias ni acciones. Cualquier predicción resultaría imposible. El progreso dejaría de ser una variable conocida (si es que alguna vez lo fue).

En otras palabras, Von Neumann sugirió a Ulam una escatología, un posible final de los tiempos. Un par de décadas más tarde, Good creyó haber encontrado el mecanismo para ello: el ordenador digital.

Vernor Vinge, científico informático de la UCLA y ganador del premio Hugo, introdujo en 1986 la «singularidad» en la informática y, más concretamente, en la inteligencia artificial con su libro de ciencia ficción Naufragio en el tiempo real.[3] En un artículo técnico posterior para la NASA, Vinge canalizó las ideas de Good:

En un plazo de treinta años dispondremos de los medios tecnológicos para crear una inteligencia superhumana. Poco después, la era de la humanidad habrá acabado ... Creo que es justo definir este acontecimiento como una singularidad. Se trata de un punto en el que debemos descartar nuestros modelos y una nueva realidad gobierna. A medida que vayamos acercándonos a ese punto, su sombra se cernirá con un tamaño cada vez mayor sobre los asuntos humanos, hasta que la noción se convierta en un lugar común. Sin embargo, cuando suceda al fin, quizá siga representando una gran sorpresa y un enigma aún mayor.[4]

Vinge, el científico informático, contó con la compañía de varios profesionales. A finales de los ochenta, Raymond Kurzweil, científico informático del MIT, futurólogo y emprendedor, se había convertido en el bulldog de la IA al expandir la idea de la singularidad en la ciencia de la cultura popular con una serie de publicaciones, comenzando con The Age of Intelligent Machines [«La era de las máquinas inteligentes»], de 1990, y su secuela de 1998, La era de las máquinas espirituales. En su best seller de 2005 se mostró aún más confiado: La Singularidad está cerca.

Kurzweil afirmaba que la innovación tecnológica, presentada en un gráfico histórico, es exponencial. La innovación se acelera, desde una visión histórica, como una función de tiempo. En otras palabras, el tiempo que transcurre entre una innovación tecnológica de primer orden y la siguiente no hace más que menguar. Por ejemplo, el papel apareció en el siglo II, y la imprenta tardó otros mil doscientos años en llegar —la imprenta de Gutenberg apareció en Europa en 1440—. Pero la computación apareció en los años cuarenta (en los años treinta, si contamos su tratamiento matemático), e internet —una innovación bastante capital— apareció menos de treinta años después. ¿Y la IA? Según la lógica de Kurzweil, la inteligencia de nivel humano en un ordenador se encuentra a pocas décadas de distancia, quizá menos—. Las curvas de crecimiento exponencial nos sorprenden.

Kurzweil etiquetó esa idea como «ley de rendimientos acelerados» (LOAR en sus siglas inglesas) y la usó como premisa en un debate cuya conclusión fue que la IA completamente humana llegaría en 2029 y a continuación, a través de una serie de procesos de carga automática hacia máquinas más inteligentes, la

superinteligencia lo haría hacia el año 2045.[5]

La superinteligencia marcaba el punto sin retorno, el lugar donde la senda del progreso desaparece en lo desconocido, en la singularidad. Se trata del punto de intersección, donde las máquinas, y no la gente, se convierten en los seres más inteligentes del planeta.

Como es sabido, Kurzweil considera que ese proceso es completamente «científico»; cita para ello la LOAR (aunque la LOAR no sea ninguna ley) y, en gran medida, se apoya en su propio entusiasmo y en sus credenciales como experto informático e inventor (Kurzweil ayudó a desarrollar la tecnología de conversión de texto a voz, lo que condujo a sistemas modernos como Siri).

Turing. Good. Vinge. Las ideas acerca de un cambio radical posibilitado por los avances informáticos ya estaban en el aire. A todas luces Kurzweil les proporcionó un mapa de carreteras. Como sucede con tantos otros sujetos obsesionados con el tema de la IA, su prosa es más papista que el papa:

Estamos entrando en una nueva era. Yo la llamo «la Singularidad». Se trata de una fusión de la inteligencia humana y la inteligencia de la máquina que va a crear algo más importante que la suma de sus partes. Se trata del proceso evolutivo más vanguardista de nuestro planeta. Se podría argumentar con firmeza que, de hecho, se trata de la vanguardia de la evolución de la inteligencia en general, porque no existen indicios de que haya sucedido en ningún otro lugar. Para mí, es la esencia misma de la civilización humana. Es algo que forma parte de nuestro destino y del destino de la evolución de cara a seguir progresando aún a mayor velocidad, y para que el poder de la inteligencia crezca de manera exponencial. Pensar en ponerle fin —pensar que los seres humanos ya estamos bien así— forma parte de un recuerdo bonito pero inapropiado de lo que fue la raza humana. La raza humana es una especie que ha experimentado una evolución cultural y tecnológica, y la naturaleza de esa evolución consiste en acelerarse, y en que su potencia crezca de manera exponencial, y de eso es de lo que estamos hablando. Su estadio siguiente consistirá en amplificar nuestra propia capacidad intelectual con los resultados de nuestra tecnología.[6]

No obstante, para ser sinceros hemos de decir que, en el momento en que

Kurzweil se subió al tren con tanto entusiasmo —de hecho, décadas antes de 1980—, el trabajo en pos de una IA científica ya había extinguido la esperanza de una marcha inexorable hacia la superinteligencia. La investigación en IA y su desarrollo se habían revelado, en una palabra, difíciles.

En los años setenta, Hubert Dreyfus, filósofo del MIT y mosca cojonera de la IA, había publicado ya una influyente refutación de esa disciplina en cuanto ejemplo clamoroso de lo que el filósofo húngaro Imre Lakatos llamó «programa de investigación degenerativo».[7] Dreyfus se puso estupendo, pero algo de razón llevaba, como los propios científicos informáticos saben demasiado bien. La disciplina sufría un revés tras otro, con aquellos intentos bien financiados y aquellas declaraciones grandilocuentes sobre máquinas inteligentes que se iban quedando cortos de manera constante (y a menudo dramática). Los laboratorios de investigación del MIT, Stanford y demás lugares se encontraron con una sucesión en apariencia interminable de disyuntivas, dificultades, confusiones y fracasos absolutos. Por ejemplo, en los años cincuenta se pensó que, dedicándole el esfuerzo de investigación y los dólares suficientes, se podría resolver el problema de la Máquina Traductora Completamente Automatizada de Alta Calidad. En los sesenta, tras una sucesión de fracasos, la inversión gubernamental en traducción había desaparecido. La esperanza de construir robots dotados de sentido común (pongamos que de la capacidad para entender el inglés y hablarlo) también se había evaporado —o al menos se había visto drásticamente reducida por aquella oleada de decepciones tempranas—. Los sistemas conversacionales que debían pasar el test de Turing de manera realista lograban burlar a los interrogadores humanos solo gracias a sus trucos y sus engaños —no a una comprensión real de la lengua—, problema que continúa haciendo mella hoy en día en los trabajos en pos de un lenguaje natural en la IA. Esta aparecía inevitablemente en las notas de prensa y en las charlas sobre el futuro, pero no sucedía lo mismo cuando la atención se centraba en el trabajo real de los laboratorios de investigación. Programar una máquina que fuera inteligente de verdad resultó ser difícil. Muy difícil.

Mientras la idea de la singularidad iba ganando fuerza en la cultura popular, los científicos de la IA seguían adentrándose en problemas de ingeniería en apariencia interminables. Y el cielo no acababa de caer. La singularidad no se acercaba. La ficción popular de Vinge seguía siendo eso: ficción.

Un examen más minucioso de la IA revela una brecha vergonzosa entre el progreso real que han realizado los científicos informáticos que se dedican a ella

y las visiones futuristas que estos y otras personas disfrutan describiendo. En 1950, Turing propuso la comprobación de una pregunta: ¿pueden las máquinas ser tan listas como las personas? Good, Vinge, Kurzweil y compañía han contestado a ella con un sí rotundo, sin tomarse en serio la verdadera naturaleza de los problemas con los que se encuentra el trabajo en ese terreno.

Esa brecha resulta muy ilustrativa.

En concreto, el fracaso de la IA a la hora de seguir un proceso sustancial sobre los aspectos más complicados de la comprensión de un lenguaje natural sugiere que las diferencias entre mente y máquina son más sutiles y complicadas de lo que Turing se imaginó. Y, si la historia de la IA nos ha de servir de guía, estas representan una dificultad profunda para la disciplina.

Nos centraremos en ellas a continuación.

Capítulo 5

La comprensión del lenguaje natural

La inteligencia artificial como disciplina oficial comenzó, con buenos auspicios, en 1956, en la ahora famosa Conferencia de Dartmouth. Entre los asistentes hubo celebridades como Shannon, de Bell Labs (teoría de la información), Marvin Minsky, de Harvard (matemáticas), Herbert Simon (destacado economista de Carnegie Mellon), John McCarthy, George Miller (psicólogo de Harvard conocido por su trabajo sobre la memoria humana) y John Nash (el matemático laureado con el Nobel al que retrató la popular película de 2001 Una mente maravillosa).

Durante la conferencia, McCarthy, que por entonces estaba en Dartmouth, pero muy pronto iba a aceptar un puesto en Stanford para desarrollar el nuevo ámbito de la ciencia informática, acuñó el término «inteligencia artificial» y dio así nombre oficial al proyecto moderno de diseñar vida inteligente. En 1816, la joven y precoz Mary Shelley había comenzado a trabajar en su obra maestra, Frankenstein. Ciento cuarenta años más tarde, los científicos reunidos en Dartmouth consideraron el montaje de un nuevo «moderno Prometeo», que no tardaría en irrumpir ante la opinión pública.

La disciplina cayó en la hipérbole desde el minuto uno. Las actas de la conferencia lo decían todo:

Proponemos que, a lo largo de dos meses del verano de 1956, diez hombres lleven a cabo un estudio sobre la inteligencia artificial en el Dartmouth College de Hanover, Nuevo Hampshire. Ese estudio debe tomar como base la conjetura de que todos los aspectos del aprendizaje o de cualquier otro rasgo de la inteligencia se pueden describir por principio con la precisión necesaria como para que una máquina pueda imitarlos. Se intentará averiguar cómo lograr que las máquinas utilicen el lenguaje, establezcan abstracciones y conceptos, resuelvan tipos de problemas que ahora están reservados a los seres humanos y

se perfeccionen a sí mismas. Creemos que se puede realizar un avance notable en uno o más de estos problemas si un grupo de científicos cuidadosamente escogido trabaja de manera conjunta en ellos durante un verano.[1]

La agenda de Dartmouth fue simple: investigar la naturaleza de las capacidades cognitivas (el pensamiento), diseñar programas que las reprodujeran e implementar y comprobar su desempeño en las nuevas computadoras electrónicas. Tal y como los participantes de Dartmouth dejaron claro, en el verano de 1956, gracias a diez investigadores armados con el conocimiento de sus respectivos campos científicos, esperaban un «avance significativo» de cara a diseñar una inteligencia humana en una máquina.

Mientras trabajaban en RAND, Herbert Simon y Allan Newell diseñaron a finales de los años cincuenta unos programas de IA que parecían cumplir las promesas optimistas de la Conferencia de Dartmouth. El programa de IA Logic Theorist, y más tarde el Solucionador General de Problemas, se sirvieron de una simple búsqueda heurística para demostrar teoremas de lógica tradicional y resolver puzles de base lógica con claros pasos computacionales. Los programas funcionaron y la IA pareció destinada a desentrañar con rapidez los secretos de la inteligencia humana, tal y como habían declarado los organizadores de Dartmouth.

Los éxitos iniciales de Simon y Newell no tardaron en animar a otros investigadores para que se marcaran objetivos más ambiciosos. Turing ya había señalado el objetivo final del programa una década antes con su versión del juego de la imitación: el test de Turing. Los científicos de Dartmouth también pensaron que programar una máquina para que entendiera el inglés o cualquier otra lengua natural representaría una declaración de victoria para la IA. Los investigadores llevaban mucho tiempo pensando que la comprensión de un lenguaje natural sería «IA completo», jerga que habían tomado prestada de las matemáticas para indicar que en el momento en que las computadoras dominaran un lenguaje natural habrían alcanzado la inteligencia general, y serían por consiguiente capaces de pensar y actuar como los seres humanos. En los años sesenta, pues, el objetivo de la IA era la traducción automática —el trasvase completo y automatizado de textos de una lengua, como podría ser el ruso, a otra, como el inglés—. La IA «iba a por todas».

LA COMPRENSIÓN DEL LENGUAJE NATURAL

Cuando la IA se centró en la comprensión del lenguaje natural, sus fieles irradiaron una seguridad en el éxito inminente de la empresa, digno de la tradición iniciada en Dartmouth. Herbert Simon, quien acabaría ganando el prestigioso premio A. M. Turing y más tarde el premio de Economía Conmemorativo de Alfred Nobel, anunció en 1957 que «ahora mismo hay en el mundo máquinas que piensan, que aprenden y que crean». En 1965 predijo que, hacia 1985, «las máquinas serán capaces de realizar cualquier trabajo que haga un hombre». Marvin Minsky también declaró en 1967 que «en el lapso de una generación, el problema de crear una "inteligencia artificial" se habrá solucionado de manera sustancial».[2]

Pero los investigadores no tardarían en descubrir que el de la traducción automática era un partido completamente diferente. Partieron del supuesto simplista de que se podía entender el lenguaje a partir del análisis de las palabras de textos amplios (llamados «corpus») utilizando técnicas estadísticas, pero pronto se demostró que estaban equivocados. Las computadoras posibilitaron la traducción automática, pero los resultados no fueron de buena calidad. Ni siquiera los programas que operaban en dominios específicos, como el de la literatura biomédica, eran inmunes al error, y esos errores eran a menudo estúpidos y vergonzosos.

Los investigadores de la traducción automática respondieron expandiendo su estrategia y explorando métodos para «diseccionar» las frases, para hallar su estructura sintáctica, usando las nuevas y potentes gramáticas «transformacionales» que había desarrollado un joven lingüista del MIT que pronto iba a obtener fama mundial: Noam Chomsky. Pero la labor misma de extraer análisis correctos de los textos de lenguaje natural se reveló muchísimo más difícil y compleja de lo que nadie se hubiera imaginado. Aparecieron problemas que, en retrospectiva, deberían haber resultado evidentes. Entre ellos se encontraba la ambigüedad en el sentido de las palabras (cuando una palabra como «banco» presenta diferentes significados posibles), la dependencia contextual deslocalizada (cuando el significado de la palabra depende de otras palabras del texto o discurso que no se encuentran próximas a ella) y otros

asuntos relacionados con la referencia (anáfora), la metáfora y la semántica (el significado). Tal y como lo definió el filósofo y científico cognitivo Jerry Fodor, la IA se había metido en una partida de ajedrez tridimensional pensando que iba a jugar al tres en raya.[3] A mediados de los años sesenta, el National Resource Council de Estados Unidos invertía millones de dólares en los trabajos sobre traducción automática de numerosas universidades del país, pero, por decirlo suavemente, los sistemas de ingeniería que comprendieran o que incluso fingieran comprender los textos de lenguaje natural con éxito eran bastante escasos.

El investigador del MIT Yehoshua Bar-Hillel, en su momento partidario feroz y entusiasta de la traducción completamente automatizada, fue el primero en dar la voz de alarma. Hizo algo más que eso, de hecho, con una serie de informes oficiales para el NRC, acompañados de unas ahora famosas notas a pie de página, en los que explicó con todo lujo de detalle los profundos problemas a los que se enfrentaba la disciplina.[4] Sus informes tuvieron un efecto sísmico sobre la comunidad investigadora. Había precisado con exactitud el obstáculo contra el que se estrellaba la traducción automática, y este era irritantemente «filosófico»: la escasez de un supuesto sentido común o «conocimiento del mundo» —el conocimiento del mundo real—. Pensemos en esta frase simple del inglés: «The box is in the pen», que la máquina podría traducir al castellano por «La caja está en el bolígrafo». Bar-Hillel explicó que esta frase servía para confundir a todos los sistemas automatizados, sin importar su grado de sofisticación, porque estos carecían de un conocimiento simple y real del mundo. El conocimiento sobre los tamaños relativos de bolígrafos y cajas capacita a los seres humanos para eliminar la ambigüedad de ese tipo de frases de manera casi instantánea. Reconocemos con rapidez que lo más probable es que el segundo elemento de la frase no sea un instrumento de escritura, sino un recinto para niños pequeños o animales, puesto que «pen» significa tanto «bolígrafo» como «redil» o «parque infantil». Y todo se vuelve más evidente con un poco de contexto adicional, como en el ejemplo de Bar-Hillel: «El pequeño John buscaba la caja de sus juguetes. Al final la encontró. La caja estaba en el parque infantil. John se puso muy contento». Pero los sistemas automatizados, que carecen de ese conocimiento, se enfrentan a una tarea misteriosa y en apariencia imposible.

Tal y como señaló Bar-Hillel, a las computadoras no se les podía suministrar ese conocimiento del mundo —al menos no de manera directa, ingenieril— porque «el número de datos que nosotros, los seres humanos, conocemos es, en un sentido creativo y fértil, infinito».[5] Sin querer había descubierto que las personas

sabemos mucho más de lo que cualquiera podría haberse imaginado —lo cual representaba lo opuesto a una solución rápida y simple para la IA—. Y eran los hechos cotidianos, de sentido común, en apariencia mundanos, acerca de la vida diaria, los que confundían a los más sofisticados sistemas automatizados. Cualquier hecho en apariencia normal podía volverse relevante en el curso de una traducción y hacer que los volúmenes de conocimiento necesario y abierto a los que podían acceder las computadoras en tiempo real o casi real, y que les infundían la capacidad cognitiva para seleccionar hechos relevantes contra ese contexto abierto (posiblemente infinito), parecieran inútiles. Tal y como concluyó Bar-Hillel en su célebre informe de 1966 para el NRC, la idea de que las computadoras pudieran ser programadas con el conocimiento del mundo de los seres humanos era «una completa quimera y difícilmente merece un debate mayor».[6]

En otras palabras, la traducción automática se había atascado con unos resultados que estaban a años luz de ser traducciones completamente automáticas de calidad elevada (y siguen siendo así, aunque la calidad haya mejorado). Por consiguiente, se mantenía el patrón. La IA se había excedido con la propaganda y, a resultas del fracaso de la investigación en traducción para estar a la altura de las promesas realizadas, el NRC retiró su financiación tras haber invertido más de veinte millones de dólares en investigación y desarrollo, una suma enorme en aquel momento. Como resultado de aquella debacle, los investigadores de IA perdieron sus trabajos, se destruyeron carreras y la IA como disciplina tuvo que volver a la casilla de salida.[7]

Durante los años setenta y ochenta, las tentativas por controlar o resolver el «problema del conocimiento del sentido común» dominaron los esfuerzos de la investigación en IA. No obstante, a principios de la década de 1990, la IA seguía sin contar con una estrategia o respuesta novedosa para su problema medular científico —y filosófico—. Japón había invertido millones en la Quinta Generación, un proyecto de perfil alto que aspiraba a alcanzar éxitos en el terreno de la robótica, y también fracasó de manera harto espectacular. A mediados de aquella década, la IA volvió a sus cuarteles de invierno —sin confianza en las promesas de sus investigadores, sin resultados para demostrar a sus detractores que estaban equivocados y sin financiación—. Entonces llegó la red.

LA RED DE EXTENSIÓN MUNDIAL

La aparición de la World Wide Web, la Red de Extensión Mundial, estimuló el resurgimiento de la IA por una simple razón: los datos. De repente, la disponibilidad de conjuntos masivos de datos, y sobre todo de corpus textuales (páginas web), debidos al esfuerzo conjunto de millones de nuevos usuarios de la red, insufló vida a las viejas y «superficiales» estrategias de estadística y reconocimiento de modelos. De repente, lo que solía ser superficial se volvió adecuado y comenzó a funcionar. Los algoritmos de aprendizaje supervisado, como las redes neuronales artificiales (redes neuronales, para abreviar), los árboles de decisión y los clasificadores bayesianos existían desde hacía décadas en los laboratorios universitarios. Pero, sin conjuntos amplios de datos, aún no habían revelado su potencial en problemas de interés como el reconocimiento facial, o la clasificación de textos, o el correo basura, o la detección de fraudes. De repente, esos métodos parecían estar llenos de promesas sin fin —entre ellas, la de crear aplicaciones rentables en el mundo real que iban a provocar una nueva ola de atención y de financiación hacia la IA.

Así nacieron los macrodatos, el big data (aunque el término llegó un poco más tarde). Con el cambio de siglo, los enfoques tipificados por algoritmos de aprendizaje como las redes neuronales y los modelos gráficos —supuestamente superficiales, ascendentes, empíricos o impulsados por datos— pasaron a amparar vastas oportunidades tanto en el terreno de la investigación en IA como en el de sus aplicaciones comerciales. Se desarrollaron nuevos métodos —incorporando modelos de Márkov ocultos, modelos de entropía máxima, campos aleatorios condicionales y clasificadores de margen amplio, como máquinas de vectores de soporte— que pasaron a dominar con rapidez la investigación pura y aplicada en la IA. En apariencia, de la noche a la mañana surgió una ciencia completa de análisis numéricos y estadísticos, basada en la optimización de los métodos de aprendizaje que trabajaban con macrodatos. Las universidades pusieron en marcha proyectos de comprensión y procesamiento del lenguaje natural. Encontraron la manera, por ejemplo, de extraer nombres y otros modelos de las páginas web (una habilidad conocida como reconocimiento de entidades), de desambiguar palabras polisémicas (con múltiples sentidos) como «banco», de realizar tareas específicas de la red como la ordenación y recuperación de páginas web (siendo su ejemplo más famoso el PageRank de Google, que Larry Page y Sergey Brin desarrollaron en la década de 1990 como estudiantes de

posgrado en Stanford), de clasificar noticias y otras páginas web por temática, de filtrar el correo basura en el correo electrónico y de ofrecer recomendaciones de productos espontáneas en sitios comerciales como Amazon. La lista no tiene fin.

El trasvase entre los enfoques lingüísticos y basados en reglas y los métodos impulsados por datos o «empíricos» pareció liberar la IA de aquellos primeros días de trabajo tormentoso en la traducción automática, cuando los problemas en apariencia interminables sobre la captura de significados y el contexto azotaron las labores de ingeniería. De hecho, la traducción automática se acabó resolviendo gracias a un grupo de investigadores de IBM que usaron un enfoque estadístico (es decir, de base no gramatical) que, en esencia, no dejaba de ser una aplicación ingeniosa de los trabajos iniciales de Claude Shannon sobre la teoría de la información. Se le conoce como el enfoque del «canal ruidoso» y considera las frases de un lenguaje de origen (pongamos que el francés) y un lenguaje de destino (pongamos que el inglés) como un intercambio de información en el que las malas traducciones constituyen una forma de ruido —lo que conduce a que la tarea del sistema consista en reducir el ruido del canal de traducción entre las frases de origen y de destino—. La idea funcionó, y las máquinas comenzaron a generar traducciones útiles sirviéndose del enfoque estadístico liderado por los laboratorios de investigación de IBM, muchísimo más simple pero con un uso intensivo de datos.

ÉXITO... O NO

El éxito de sistemas contemporáneos como Google Translate ante el problema antaño desconcertante de la traducción automática se publicita a menudo como una prueba de que la IA acabará triunfando si se le proporcionan el tiempo suficiente y las ideas adecuadas. La realidad invita a poner los pies en el suelo.

Por un lado, resulta que ciertos problemas sobre la comprensión de un lenguaje natural se pueden abordar con enfoques estadísticos o de aprendizaje automático; pero, por otro, las preocupaciones originales de Bar-Hillel y otros acerca de la semántica (el significado) y la pragmática (el contexto) se han revelado fundamentadas. La traducción automática, que parecía representar un problema relacionado con el lenguaje natural de muy difícil solución, pudo realizarse de

manera adecuada gracias a análisis estadísticos simples que contaran con corpus amplios (conjuntos de datos) en lenguajes diferentes. (Y asumamos que la traducción automática aún no presenta una calidad demasiado elevada; vendría a ser solo «razonablemente buena».) Esto no demuestra que la inteligencia de las máquinas haya crecido de manera impactante a la hora de entender el lenguaje natural, sino que la traducción automática es un problema mucho más sencillo de lo que se pensó en un primer momento.

Una vez más, siguen existiendo grandes problemas para que los ordenadores comprendan el lenguaje. Una manera simple de verlo consiste en recuperar el test de Turing y replanteárselo a la luz de la historia de la IA, con numerosos y en su mayoría infructuosos esfuerzos para resolver los problemas que la acompañan, o incluso para realizar algún avance sustancial. Lo más probable es que futurólogos como Nick Bostrom, junto con la mayor parte de la comunidad científica de la IA, deseen que el público se olvide del test.

No se trata —como se dice a veces— de que el test sea fallido o sirva de poco. Simplemente es demasiado difícil.

EL TEST DE TURING

Si se las observaba a diez mil metros de distancia, en efecto pudo parecer que las computadoras iban ganando inteligencia mientras la IA progresaba desde su génesis con los primeros trabajos de Turing y el inicio de la conferencia de Dartmouth. Sin duda, las computadoras pasaron a tener procesadores más potentes y una mayor capacidad de almacenaje, y eso condujo de manera inevitable al desarrollo de unos problemas bien definidos. A la hora de jugar al ajedrez, las computadoras no sobrepasaban la capacidad de las personas más cualificadas, pero tampoco dejaban de mejorar. Programas como DENDRAL, que analizaba estructuras químicas, y MYCIN, que a veces ofrecía diagnósticos médicos bastante buenos, dejaron a las claras que los métodos de la IA eran relevantes para diversos problemas que por lo general requerirían de una inteligencia humana elevada. Tal y como ya hemos visto, en un primer momento la traducción automática fue un fracaso, pero comenzó a ofrecer rendimiento con diferentes enfoques posibilitados por la disponibilidad de grandes conjuntos de

datos (fue la antesala de numerosos éxitos con los macrodatos y la ciencia de datos en los 2000). Los sistemas de IA, cada vez más potentes y sofisticados, comenzaron a realizar una labor de zapa con todo tipo de tareas de procesamiento de lenguaje natural, como la de generar análisis sintácticos de oraciones de lenguaje natural y etiquetar partes del discurso o entidades (personas, organizaciones, lugares y cosas así).[8]

Sin embargo, seguía sin alcanzarse el objetivo original de Turing para la IA: que lograra pasar su famoso test. Para la IA, comprender una conversación no era una tarea bien definida; se trataba de toda una aptitud. Aunque existió alguna esperanza de éxito con sistemas tempranos como el SHRDLU de Terry Winograd, que apilaba y manipulaba bloques con un brazo robótico a través del análisis sintáctico de órdenes como «Coloca el bloque A encima del B, y coloca ambos encima del C», los investigadores no llegaban a nada con los sistemas conversacionales abiertos. Los avances se basaban en engaños, en estratagemas ingeniosas de los diseñadores de sistemas para burlar a los interlocutores humanos fingiendo ignorancia o adoptando respuestas evasivas y tácticas diversas cuando las preguntas sacaban al programa de su elemento. El programa ELIZA de Joseph Weizenbaum, quizá el sistema trucado de mayor fama, podía generar la ilusión de un diálogo humano imitando los comentarios de un psicoterapeuta rogeriano a alguno de sus pacientes. He aquí una muestra, con las respuestas de ELIZA en cursiva:

—Todos los hombres son iguales.

—En qué sentido

—Siempre nos están chinchando por un motivo u otro.

—Se te ocurre un ejemplo específico de lo que dices

—Bueno, mi novio me ha hecho venir aquí.

—Tu novio te ha hecho venir aquí

—Dice que estoy deprimida la mayor parte del tiempo.

—Lamento oír que estás deprimida

—Es cierto. No soy feliz.

—Crees que venir aquí te ayudará a ser más feliz

—Necesito ayuda, al menos eso parece ser cierto.[9]

Pero esos programas, según los parámetros de la verdadera IA, eran fraudes evidentes.

Como también lo fue el programa que en múltiples ocasiones burló momentáneamente a los jueces del premio Loebner, una puesta en escena del test de Turing con carácter anual. La prensa se apresuró a declarar que al fin se había superado el test, y siguió una breve pero intensa celebración a bombo y platillo del avance de la IA ante su más famoso desafío. El bot conversacional —llamado Eugene Goostman— se presentó como un muchacho ucraniano de trece años, identidad diseñada para que los jueces anticiparan (y excusaran) sus errores gramaticales y sus lagunas de conocimientos generales. El bot también contestó a algunas preguntas no con respuestas directas, sino con distracciones y evasivas en forma de comentarios aleatorios y chascarrillos digresivos. ¿Engañó el bot por momentos a algunos de los jueces? Sí. ¿Superó en realidad el test de Turing de manera razonable? No.[10]

Aunque durante estos últimos años se ha podido dominar (en cierto modo) la traducción automática gracias a los amplios volúmenes de textos traducidos a diferentes idiomas que hay en la web, el test de Turing sigue representando una frustración perpetua para la IA. El fantasma de Bar-Hillel continúa acosándonos.

Capítulo 6

De la IA como tecnología kitsch

En 1980, el escritor de origen checo Milan Kundera escribió su obra maestra, La insoportable levedad del ser. Se trata de una historia de amor con el telón de fondo de la invasión de Checoslovaquia en 1968 por parte de la Unión Soviética. Kundera escribió acerca de los escritores y artistas que se suicidaron tras ser acosados con calumnias de forma incansable por la policía secreta soviética, cuyos elementos se habían insertado en el tejido social, intelectual y cultural de Praga. Muertos y desacreditados, los intelectuales praguenses sufrían a continuación una vergüenza más (aunque póstuma): el elogio repugnante, durante sus funerales, por parte de miembros y oficiales del partido soviético, que daban fe de la devoción por el Estado que el finado había sentido a lo largo de toda su vida. La propaganda soviética los conducía a la muerte y, acto seguido, esa misma propaganda presentaba sus existencias como si las hubieran sacrificado con nobleza para promover unas ideas contra las que, de hecho, se habían manifestado tanto en público como en privado. Se decía que habían amado aquello que más odiaban.

La propaganda soviética era implacable, pero no era ni furiosa ni estúpida. Tenía un propósito concreto, y ese propósito consistía en purgar el país de aquellas expresiones profundas y elevadas (y oposicionistas) acerca del sentido del país, de su gente y de su vida. Los soviéticos purgaron Praga, y Checoslovaquia entera, de su historia compartida, de sus tradiciones y de su sentido de lo que resultaba precioso, aquello por lo que valdría la pena luchar. Tras silenciar a los librepensadores, como quien pinta una pared después de pulirla, los soviéticos serían libres para imponer su visión del mundo sin enfrentarse a una oposición seria u organizada. El relato de Kundera es un recuento incisivo y a menudo trágico sobre el valor de la vida humana y la manera en que una creencia o ideología en particular puede intentar —pero nunca conseguir por completo— ofuscar y encubrir todo aquello que resulta importante para el individuo y la sociedad. Kundera definió la cultura que los soviéticos endosaron al pueblo checo derrotado como kitsch.

LA TECNOLOGÍA KITSCH

«Kitsch» es una palabra alemana que, si bien hoy en día se aplica a obras de arte o motivos decorativos cursis o chabacanos, en su origen implicaba un sentimentalismo y melodrama exagerados en cualquier ámbito. Los errores de inteligencia en el núcleo de la cosmovisión de la IA —sus creencias; es decir, no la ciencia— han dado pie a una forma moderna y particularmente perniciosa de kitsch. El sueño de un superordenador inteligente no es como la propaganda soviética, y nadie nos obliga a creer en el advenimiento de las máquinas. Pero ambos sí comparten la idea básica de reemplazar debates difíciles y complejos sobre el individuo y la sociedad con historias tecnológicas que, al igual que la cultura soviética, reescriben ideas viejas con abstracciones peligrosamente unidimensionales.

El significado y el uso de la palabra kitsch han cambiado con el paso del tiempo. La definición original en alemán difiere de algún modo del significado que pretendo explorar aquí, pero hay dos ingredientes esenciales de ese sentido inicial que deberían ofrecer la suficiente claridad a mi afirmación. Primero, kitsch implica la simplificación de unas ideas complicadas. Tiene que haber una historia simple que se pueda contar. Segundo, ofrece soluciones sencillas que barren con la emoción las preguntas y la confusión que la gente experimenta hacia sus problemas vitales, en vez de abordar esas preguntas con un debate serio y exploratorio. Así, encontramos un ejemplo perfecto de kitsch en la ensoñación de que algún día un androide impresionante y dotado de superinteligencia remodelará la sociedad humana y sus viejas tradiciones e ideas, y que entraremos en una nueva era, por fortuna libre de las antiguas discusiones acerca de Dios, la mente, la libertad, la buena vida y demás. Unas máquinas hermosas (o unas máquinas dotadas de una bella inteligencia), como la Ava de la película de ciencia ficción de 2015 Ex Machina, a la que dio vida Alicia Vikander, acabarán con la dura realidad de la existencia humana. Eso es kitsch al estilo tecnológico. Como la propaganda soviética, puede horrorizarnos o apaciguarnos, pero nos proporciona una historia nueva que se escribe sobre la anterior y que vuelve innecesario lo que era verdad antes, de modo que la antigua realidad desaparece.

Pese a todas sus contribuciones a la ciencia y la ingeniería, Alan Turing hizo posible la génesis y el crecimiento viral del kitsch en la tecnología al comenzar equiparando la inteligencia con la resolución de errores. Más adelante, Jack Good agravó el error de la inteligencia de Turing con su muy discutida idea de ultrainteligencia y la propuesta de que la aparición de las máquinas inteligentes implicaría de manera necesaria la llegada de las máquinas superinteligentes. Una vez que la imaginación popular hubo aceptado la idea de las máquinas superinteligentes, la reescritura del sentido de la existencia humana, su significado e historia se podría efectuar dentro de los parámetros de la informática y de la tecnología.

Pero las máquinas ultrainteligentes son una fantasía, y pretender lo contrario da aliento a ese proceso indeseable que conduce a la tecnología kitsch, por lo general de dos maneras igual de superficiales. En un extremo escuchamos un relato de IA apocalíptico o aterrador, como una de esas historias que se cuentan alrededor del fuego en un campamento. En el otro extremo nos encontramos con la IA utópica o soñadora, que es igual de superficial e inmerecida. Si nos tomamos en serio una u otra forma de la IA kitsch, acabaremos en un mundo definido tan solo por la tecnología.

Es un tema que retomaré más tarde, porque expone el problema fundamental de la IA futurista. Tal y como dice Nathan, el genial científico informático de Ex Machina: «Llegará el día en que las inteligencias artificiales vuelvan la vista atrás y nos vean tal y como nosotros vemos los esqueletos fósiles de las llanuras africanas. Un simio erecto que vivió rodeado de polvo, con una lengua en bruto y herramientas sin pulir, listo para extinguirse». La verdad es que no está claro que vaya a haber algún ordenador capaz de volver la vista atrás. Ese sentimiento popular requiere de una inmersión profunda en el sentido de la existencia, de la vida, de la consciencia y de la inteligencia, y en las diferencias que existen entre nosotros y la informática y sus numerosas tecnologías. Lo kitsch evita que tratemos de resolver la naturaleza humana y demás empeños filosóficos de gravedad. Y no debería ser así, como bien sabía Kundera.

Lo kitsch hunde sus raíces, por lo general, en un sistema de pensamiento más amplio. Para los comunistas, este fue el marxismo. En el mito de la inevitabilidad, es la tecnociencia. Nuestra visión tecnocientífica del mundo es una herencia directa de la obra de Auguste Comte.

NUESTRA CONDICIÓN TECNOCIENTÍFICA

Probablemente, el primer pensador que explicó y desarrolló de manera completa la cosmovisión de la tecnociencia fue Auguste Comte.[1] Comte, filósofo del siglo XIX reconocido popularmente por haber sentado los fundamentos de la sociología como un campo de estudio científico, desarrolló y expuso la teoría del positivismo. Según esa visión, los únicos fenómenos que existen son aquellos observables y científicos; la religión y la filosofía serían imaginarios. Comte explicitó, primero, su idea de que la mente humana se dirige hacia la verdad, igual que la sociedad en su conjunto, a través de unas etapas que se inician con los pensamientos religioso y filosófico, y que progresan hacia el pensamiento científico. Y, segundo, explicó que la tecnociencia acabaría por crear un cielo en la tierra al posibilitar la comprensión de la naturaleza de todas las cosas (ciencia), y al servirse a continuación de ese conocimiento para desarrollar tecnologías que alarguen de manera notable nuestras vidas, que las mejoren y que las vuelvan más valiosas.

La explicación de Comte sobre el poder transformativo de la tecnociencia acabó dando pie a la convicción de que la religión y la Iglesia en particular podrían ser reemplazadas por una «religión de la humanidad» plenamente secular, que no creería en ningún dios y que estaría anclada con firmeza en las ciencias y en la realidad material. En el momento en que Comte escribió todo eso, durante el siglo XIX, había suficientes pruebas sobre el poder del pensamiento humano tanto para descubrir leyes científicas como para crear tecnologías útiles y potentes, de modo que la tecnociencia arraigó en el centro mismo de la mente moderna.

No obstante, la teoría de Comte provocó recelos desde un primer momento. Nietzsche, por ejemplo, lamentó que la idea de persona se viera restringida y limitada de esa manera. La tecnociencia podía ayudarnos a vivir más tiempo, pero no podía hacernos más sabios. La idea del héroe, o de una persona con dones y virtudes extraordinarios y merecidos, no entraba en la visión de Comte, que en esencia había reemplazado el debate tradicional acerca de la persona con el debate sobre el progreso de la ciencia y, en especial, de la tecnología.[2]

El materialismo de Comte también sugirió a otros pensadores una disminución, en vez de la expansión, de las posibilidades humanas. Desde el este, en Rusia, el

escritor Dostoyevski protestó desdeñoso contra el «flagelo» creciente de la creencia absoluta en el materialismo y el cientificismo —la visión según la cual el científico es el único conocimiento real—, y lo hizo con una prosa que reflejaba el escepticismo de otros pensadores e incluso su miedo ante la velocidad con la que iba ganando predominancia el pensamiento tecnocientífico. Tal y como dijo en Memorias del subsuelo: «Nuestro propio deseo, voluntario y libre; nuestro propio capricho, aun el más alocado; la fantasía desatada hasta rayar en lo extravagante...; he ahí en qué consiste la ventaja pasada por alto, el interés más principal, que en ninguna clasificación se incluye y que manda a paseo todos los sistemas y teorías».*a [3]

Dostoyevski, Nietzsche y otros apuntaban al ideal de la persona plena, pero Comte hablaba de un ideal externo a nosotros: la tecnociencia y su progreso. El problema consistía en que, como bien sabía Comte, la visión de un futuro tecnocientífico representaba también una declaración profunda y cargada de sentido sobre la naturaleza de la persona. En efecto, Comte afirmó que las concepciones tradicionales de la persona —como algo único porque lo creó Dios, o como alguien en busca de la sabiduría (no solo del conocimiento tecnológico), tal y como sostenían los filósofos griegos— habían pasado a ser irrelevantes en virtud de los éxitos científico y tecnológico. Su filosofía tecnocientífica fue una apostilla a la esencia y las posibilidades de la naturaleza humana. Era una postura radical, y los pensadores iconoclastas, que no se dejaron engañar por el monstruo de la tecnociencia, acertaron al desafiar las ideas propuestas por Comte (y compañía).[4]

EL TRIUNFO DEL HOMO FABER

La tecnociencia triunfó durante el siglo XX, pero también siguió generando respuestas escépticas. Hannah Arendt, la filósofa que se hizo famosa por su frase sobre «la banalidad del mal», en referencia a los juicios nazis de Núremberg, afirmó que la tecnociencia de Comte —que a mediados del siglo XX ciertamente no había perdido fuerza como idea filosófica— equivalía nada más y nada menos que a una redefinición de la naturaleza humana misma.[5] Arendt apuntó hacia la comprensión clásica de los seres humanos como Homo sapiens —literalmente, «hombre sabio»— y al foco histórico sobre la sabiduría y el

conocimiento en vez de las aptitudes técnicas, y argumentó que abrazar la tecnociencia como cosmovisión implicaba redefinirnos como Homo faber —«hombre constructor».

En términos griegos, el Homo faber es una persona que cree que la téchne —el conocimiento de un oficio o acerca de la construcción de algo, la raíz de la tecnología— define aquello que somos. Esa visión faberiana de la naturaleza humana se ajusta a la perfección no solo a la idea decimonónica de Comte sobre una tecnociencia utópica, sino a la obsesión del siglo XX por construir tecnologías cada vez más potentes y que culminó con el proyecto grandioso de, en efecto, construirnos a nosotros mismos: la inteligencia artificial. Ese proyecto no tendría sentido si las nociones tradicionales sobre el significado de la humanidad hubieran permanecido intactas.

Arendt sostuvo que el cambio sísmico desde la sabiduría y el conocimiento, y hacia la tecnología y la construcción, implicaba una comprensión de nosotros mismos limitante y peligrosa en potencia, que garantizaría no solo que el desarrollo tecnológico prosiguiera desenfrenado, sino que viéramos cada vez más los éxitos tecnológicos como declaraciones valiosas acerca de nosotros mismos. En otras palabras, estábamos reduciendo nuestra propia valía a fin de aumentar, más allá de lo sabio o razonable, nuestra estimación hacia las maravillas que se podían construir con las herramientas de la tecnociencia.

Los comentarios en un principio crípticos de Von Neumann sobre esa aceleración del avance tecnológico que nos acerca a la «singularidad» ganan claridad a la luz de la posición de su contemporánea Arendt. Aunque Von Neumann, científico y matemático, no ahondó (hasta donde sabemos) en esas afirmaciones, estas reflejan a la perfección la insistencia de Arendt en la significación profunda de la tecnociencia para nosotros y para nuestro futuro —para lo que los filósofos de la tecnología llaman «la condición humana»—. A Comte quizá le pareciera perverso que la tecnología pudiera acelerarse al extremo de escapar a nuestro control, pero en ninguno de sus textos se puede encontrar el menor indicio de lo que iba a apuntar Arendt (y otros); que, al abogar por la tecnociencia como respuesta humana a los problemas humanos, nos estamos involucrando en el proyecto de redefinir nuestra comprensión de nosotros mismos. El giro hacia la téchne en vez de, pongamos, la episteme (conocimiento de los fenómenos naturales) o la sapientiae (la sabiduría relacionada con los valores humanos y sociales) dificulta la forja de una idea valiosa de la unicidad humana. (Al fin y al cabo, incluso las abejas son

constructoras; de colmenas, en su caso.)

Situar la téchne en el centro del debate también posibilita la visión de la persona como algo que se puede construir, ya que implica que su máxima consecución como tal es una capacidad superior para fabricar tecnologías cada vez más avanzadas. Una vez emprendida esa ruta, el trayecto hacia la inteligencia artificial es corto. Y ahí aparece la conexión evidente con los errores de la inteligencia que cometió primero Turing y que se hicieron extensivos a Jack Good y compañía hasta llegar al día de hoy: el triunfo definitivo del Homo faber como especie consiste en construirse a sí mismo. Y ese es precisamente, por supuesto, el objetivo declarado de la IA. Explorar la posibilidad de que el proyecto sea un éxito o no requerirá que nos sumerjamos en las aguas profundas de la comprensión de nuestra propia naturaleza.

PARA COMPLETAR EL PUZLE

La tecnociencia se inició con la revolución científica, y pocos siglos después la mayor parte de la teoría científica moderna ya había ocupado su lugar. Con raras excepciones —siendo una de ellas, evidentemente, el desarrollo de la teoría cuántica y de la relatividad durante el siglo XX—, el conocimiento científico avanzó a medida que se ponían en marcha la mayoría de las teorías de la física. El conocimiento científico era como un puzle de piezas teóricas que conformaban una imagen del mundo y del universo. La física de Newton, la electrodinámica de Maxwell, las teorías del trabajo y la termodinámica de Carnot y otros..., todas esas formas de conocimiento científico encajan entre sí y ofrecen una imagen unificada del mundo. La teoría de la evolución de Darwin, en la década de 1850, y los sucesivos descubrimientos geográficos y arqueológicos añadieron nuevas hipótesis y detalles. (Por supuesto, esas teorías eran debatidas y comprobadas, y algunas se revelaban erróneas o se revisaban.) El alcance de las posibilidades de la teoría científica, pues, iba menguando de manera extraña —como cuando uno trabaja con un rompecabezas, y con cada pieza que encaja en su lugar el número de elecciones que quedan se va volviendo cada vez más limitado.

En contraste, la innovación tecnológica estalló. Tal y como ha señalado Ray

Kurzweil, la innovación tecnológica se acelera. Una invención no limita aquello que puede venir a continuación, sino que posibilita que haya cada vez más invenciones. En otras palabras, la tecnología parece evolucionar. No la ordenamos igual que la teoría. En cambio, hacemos crecer los desarrollos tecnológicos apilándolos unos encima de los otros de manera en apariencia interminable. La aceleración de la evolución tecnológica implica simplemente que el plazo de tiempo entre una innovación tecnológica capital y la siguiente no deja de menguar, históricamente, así que la distancia entre, pongamos, la invención de la imprenta y la llegada de la computadora es muy amplia en comparación con, pongamos de nuevo, la distancia entre la computadora e internet. La fusión entre ciencia y tecnología resulta, pues, complicada, y la palabra misma «tecnociencia» sugiere que, a medida que las cosas vayan progresando, la ciencia se acomodará y la tecnología continuará evolucionando —y lo hará, tal y como dice Kurzweil, de manera exponencial.

Así que el término mismo «tecnociencia» demuestra la complejidad y el carácter impredecible de nuestro mundo. No todas las áreas de la actividad humana siguen los mismos parámetros de crecimiento; ninguna de ellas puede tenderse a lo largo de otra, como si se tratara de una plantilla. Que la inteligencia humana se parezca a la inteligencia de la máquina —o no— es algo que aún está por verse. La cuestión de la IA debería ser una invitación no a ignorar los problemas filosóficos, sino a pelearse con ellos. Y la tecnociencia, entendida como una afirmación sobre nosotros mismos, es en definitiva una simplificación terrible, que representa (entre otras cosas) la introducción de lo kitsch en la corriente de los asuntos complejos y difíciles de esta vida.

Capítulo 7

Simplificaciones y misterios

Poco antes de que Turing publicara en 1950 su «Maquinaria computacional e inteligencia», el psicólogo conductista B. F. Skinner publicó la novela de ciencia ficción Walden dos.[1] En ella, Skinner hace que sus personajes afirmen que el libre albedrío es una ilusión, y que el comportamiento de las personas se puede controlar de manera externa, desde su entorno. Si alguien (pongamos que un científico) altera el entorno, el comportamiento de esa persona en ese entorno se alterará también.

En un sentido trivial, eso es verdad. Si un déspota le niega comida, seguridad y oportunidades de trabajo a la gente, esa gente se volverá infeliz. Podemos predecir ese tipo de cambios. Lo que Skinner quería decir, en cualquier caso, es que la persona se encuentra completamente determinada por impulsos externos (por estímulos, en sus propias palabras).

De hecho, la noción skinneriana de la persona como una «caja negra» es la misma idea básica que Turing tenía en la cabeza. En las cajas negras, tratamos la salida del sistema como una función de su entrada —el funcionamiento de su interior no se llega a describir—. Skinner afirmó en Walden dos que se podría construir un mundo perfecto —una utopía— tratando a la gente como si fueran cajas negras, proporcionándoles ciertos impulsos físicos (estímulos) para obtener ciertos resultados (respuestas). Mientras tanto, Turing especulaba con la idea de que el ser humano era equivalente operativamente a una máquina compleja, y para demostrarlo sugirió construir una máquina, proporcionarle impulsos y examinar sus respuestas.

Por desgracia, esa manera de pensar prescindía de muchas cuestiones de importancia, y hoy parece evidente que hemos heredado sus errores. Mientras que la teoría del condicionamiento operante de Skinner —o conductismo, como se le llamó después— iba a ser motivo de gran controversia más tarde, durante el siglo XX, la interdisciplinaria «revolución cognitiva» que la reemplazó pasó a tratar la inteligencia como un conjunto de meros cálculos internos. Esa idea,

apuntalada por una filosofía denominada «teoría computacional de la mente», que afirma que la mente humana es un sistema de procesamiento de información, no deja de respaldar la confianza teórica en el triunfo futuro de la IA.

Aquí lo mejor es ser claros: equiparar la mente humana con un ordenador no es una actitud científica, sino filosófica.

EL DISPARATE DE LA PREDICCIÓN

Tal y como señala Stuart Russell, en la búsqueda de la inteligencia artificial no deberíamos apostar en contra del «ingenio humano».[2] Pero, de manera similar, no deberíamos realizar predicciones optimistas (o catastrofistas) sin una sólida base científica.

A los expertos e incluso (o en especial) a los científicos les encanta hacer predicciones, pero la mayoría de ellas son erróneas. En su excelente libro Future Babble, Dan Gardner documenta la tasa de acierto de las predicciones realizadas en los ámbitos de la historia y la geopolítica, hasta llegar a las ciencias.[3] Y descubre que los teóricos —expertos con grandes visiones de futuro basadas en la teoría particular que ellos mismos respaldan— tienden a realizar predicciones peores que la gente pragmática, que ve el mundo como algo complejo, que no acaba de encajar con ninguna teoría en concreto.

Gardner se refiere a la clase de los expertos y a los pensadores pragmáticos como «erizos» y «zorros» respectivamente (toma prestados los términos del psicólogo Philip Tetlock, quien a su vez le había copiado la terminología a Isaiah Berlin). De la misma manera que el erizo escarba su madriguera, los expertos erizos escarban en una idea. De forma inevitable llegan a pensar que esa idea plasma la esencia de todo, y esa fe alimenta el proselitismo consiguiente. Marx fue un erizo infatigable.

Los zorros ven la complejidad y el carácter incalculable de los asuntos del mundo, y o bien evitan realizar predicciones audaces o predicen con mayor seguridad (y quizá inteligencia) que las cosas no cambiarán de la manera que

pensamos. Para el zorro, el negocio de la predicción es casi una temeridad, porque la verdad es que no podemos saber lo que surgirá de la compleja dinámica que existe entre la geopolítica, la política interna de un estado (pongamos, ¿quién ganará unas elecciones?), la ciencia y la tecnología. Tal y como nos advirtió Lev Tolstói, novelista del siglo XIX, las guerras tienen lugar por motivos que no podemos encajar en los planes de batalla.

Algunos científicos de IA son famosos por haberse mostrado vulpinos en sus predicciones sobre esa disciplina. Mira a Yoshua Bengio, profesor de ciencias de la informática en la universidad de Montreal, Canadá, y uno de los pioneros del aprendizaje profundo: «No soy yo quien pueda decirlo —responde a la pregunta de si podemos esperar una IA de nivel humano—. No tiene sentido ni sirve de nada adivinar una fecha porque no tenemos ni idea. Lo único que puedo decir es que no sucederá en un plazo de pocos años».[4]

Ray Kurzweil ofrece una respuesta más propia de un erizo: la IA de nivel humano llegará en 2029. Invoca la «ley» de rendimientos acelerados para hacer que sus predicciones parezcan científicas y no deja de encontrar pruebas de que ha acertado con todos los supuestos avances que han tenido lugar hasta la fecha.[5]

A veces, los filósofos tienen la virtud de pensar con claridad sobre un problema porque no les estorba ningún celo concreto que pueda ligarse a los profesionales de ese terreno (los que aún deseen filosofar). Por ejemplo, Alasdair MacIntyre, en su ya clásico Tras la virtud, señaló las cuatro fuentes fundamentales de impredecibilidad que hay en el mundo. En concreto, su exposición sobre la «innovación conceptual radical» tiene una relevancia directa en las preguntas acerca de la llegada de una IA de nivel humano. MacIntyre recuerda el argumento contra la posibilidad de predecir inventos que realizó Karl Popper, filósofo de la ciencia del siglo XX:

En algún momento del Paleolítico, tú y yo estamos hablando del futuro y yo predigo que antes de que transcurran diez años alguien habrá inventado la rueda.

—¿La rueda? —preguntas—. ¿Y eso qué es?

Procedo a describirte la rueda, buscando las palabras, sin duda con dificultad, que por primera vez sirvan para decir lo que serán el aro, los radios, el buje y quizá el eje. Acto seguido hago una pausa, traspuesto.

—Pero es que nadie podrá inventar la rueda, porque la acabo de inventar yo.

En otras palabras, la invención de la rueda no se puede predecir. Y es que una parte necesaria a la hora de predecir un invento consiste en decir lo que la rueda es, y decir lo que la rueda es implica inventarla. Es sencillo apreciar la manera en que este ejemplo puede generalizarse. Cualquier invento, cualquier descubrimiento cuya esencia consista en elaborar un concepto radicalmente nuevo, no se puede predecir, ya que una parte necesaria de esa predicción es la elaboración actual del concepto mismo cuyo descubrimiento o invención solo debía tener lugar en el futuro. La idea de predecir una innovación conceptual radical es en sí misma una incoherencia conceptual.[6]

En otras palabras, sugerir que nos encontramos en la «senda» de una inteligencia artificial general cuya llegada se pueda predecir presupone que no hay ninguna innovación conceptual de camino, una postura con la que ni siquiera se mostrarían de acuerdo los científicos de IA más convencidos del advenimiento de la inteligencia artificial general y que estén dispuestos a realizar predicciones, como Ray Kurzweil. Como mínimo, todos sabemos que para que un supuesto sistema de inteligencia artificial general alcance una capacidad todavía desconocida para comprender el lenguaje natural, antes deberá darse el invento o el descubrimiento de un componente racional y generalizador. Esto cuenta sin duda a modo de ejemplo de «innovación conceptual radical», porque aún no tenemos ni idea de lo que será, ni de su aspecto siquiera.

La idea de que podemos predecir la llegada de la IA esconde por lo general una premisa, en grado diverso de aceptación, según la cual el éxito en los sistemas de IA restrictivos, como el de los juegos, facilitará una ampliación que lleve a la inteligencia general, de modo que se puede trazar con cierta confianza la línea predictiva que conducirá de la inteligencia artificial a la inteligencia artificial general. Se trata de una mala suposición, tanto para alentar los avances en el terreno hacia la inteligencia artificial general como para la lógica de los argumentos en favor de una posible predicción.

Las predicciones acerca de descubrimientos científicos quizá se puedan entender mejor como una forma de complacencia mitológica; en efecto, la certeza acerca de la llegada de la inteligencia artificial general solo puede morar en el reino de lo mítico, sin las trabas que le suponen las dudas de Popper o de MacIntyre o de

cualquier otra persona.

No toda la mitología sobre la IA es mala. Esta mantiene con vida el anhelo arquetípico por crear vida e inteligencia, y puede abrir ventanas hacia la comprensión de nosotros mismos. Pero, cuando se disfraza de ciencia y de certeza, el mito confunde al público y frustra a los investigadores fuera del ámbito mitológico, conscientes de que aún hay que resolver grandes obstáculos teóricos. «No tenemos ni idea», en palabras de Bengio. Aunque cuente con el apoyo de todas las pruebas y sea cierto, esto resulta extremada y deprimentemente pesimista para los mitólogos.

Sin embargo, los obstáculos no son siempre infranqueables, e, incluso cuando lo son —cuando nos vemos obligados a reconocer ciertos límites—, a continuación quedamos liberados para encontrar una manera diferente de alcanzar nuestro objetivo, o, según el impulso, para formular unos objetivos completamente nuevos. La historia de la ciencia está llena a reventar de ejemplos en los que el descubrimiento de un punto muerto condujo a un nuevo avance. Werner Heisenberg descubrió el principio de incertidumbre mientras intentaba solucionar las consecuencias de la nueva física cuántica. El principio sostiene que es imposible aislar la posición y la cantidad de movimiento de una partícula subatómica de manera simultánea. Eso plantea unos límites fundamentales sobre nuestra capacidad para predecir los movimientos individuales de las partículas en el campo subatómico (porque «ver» la posición de la partícula requiere hacerla chocar con un fotón, lo que también tiene el efecto de desviarla). Aunque no sea más que una limitación, el principio de incertidumbre se ha revelado tan valioso como fructífero a la hora de comprender la mecánica cuántica. Por ejemplo, no habríamos podido aspirar a construir un ordenador cuántico si antes no hubiéramos comprendido la naturaleza de la incertidumbre.

Hay muchos más ejemplos. El móvil perpetuo fue una obsesión de los siglos XVIII y XIX; su órbita atrajo a las mejores y más brillantes mentes. Los avances en las teorías del trabajo y de la termodinámica jubilaron ese sueño..., pero, de paso, ampararon un progreso inmenso en la comprensión de la energía y el movimiento. Aceptar la complejidad —y las complicaciones— nos conduce más lejos que las simplificaciones excesivas.

UN EXTRAÑO (PERO OPORTUNO) ARGUMENTO POR

PARTE DE MICHAEL POLANYI

Una posibilidad que ha surgido en el debate sobre la IA es que alcancemos la inteligencia general pero no podamos anotarla —es decir, programarla— porque en muchas áreas importantes será una caja negra para nosotros. Esto nos lleva a hablar de Michael Polanyi.

Polanyi, químico y filósofo influyente en su día pero ahora poco conocido, afirmó a mediados del siglo XX que los símbolos con los que la anotamos solo capturan una parte de la inteligencia —los usos del lenguaje que él llamó «articulaciones»—. Polanyi anticipó muchos de los dolores de cabeza que los sistemas de IA han causado a sus diseñadores; de hecho, en sus trabajos tardíos rechazó explícitamente que las máquinas pudieran capturar la totalidad de la inteligencia humana por razones derivadas del carácter incompleto de las articulaciones.

Polanyi sostuvo que las articulaciones obvian por necesidad componentes «tácitos» de la inteligencia —aspectos del pensamiento que no se pueden describir con precisión anotando símbolos—.[7] (La red neuronal que construimos también es un sistema de símbolos.) Por ejemplo, eso sirve para explicar que ciertas habilidades o artesanías, como la gastronómica, no se puedan dominar simplemente tras leer unas recetas. Hacemos cosas, pero eso no significa que podamos programar todo lo que hacemos: piensa en escribir un programa que escriba una novela del orden del Ulises de James Joyce. Ese programa carecería de sentido. En su lugar, escribiríamos la novela directamente (si fuéramos Joyce).

Los escritos de Polanyi llegaron en un mal momento para sugerir una visión opuesta a la IA, ya que la disciplina había despegado en los años cincuenta a bombo y platillo. Su defensa del conocimiento tácito fue recogida más adelante por Hubert Dreyfus con su ya comentado ataque contra la IA; quizá a causa de su tono, en ocasiones demasiado tendencioso, los comentarios de Dreyfus se convirtieron en un pararrayos de réplicas y, al menos en un primer momento, no se ganaron a los pensadores más populares de la IA. (Por desgracia, también declaró que un sistema de IA nunca podría vencer a un gran campeón de ajedrez.)[8]

Pero la posibilidad de que no todo lo que sabemos se pueda anotar representa un problema recurrente para la IA, ya que implica que los programadores están buscando la cuadratura del círculo. Estos escriben programas específicos (o programas de análisis de datos, lo cual sigue siendo específico) a los que se les escapa algo relacionado con nuestra mente. Las ideas de Polanyi sugieren que la mente y la máquina presentan diferencias fundamentales y que equiparar la una con la otra conduce a una simplificación de nuestras ideas sobre la mente. Si la mente —o al menos la inteligencia general— debe ser tratada como algo codificable o anotable, tenemos que simplificar la «mente» misma para que buena parte del debate contemporáneo cobre sentido.

EL REGRESO DE LOS ZORROS

A principios de la década de 2000, la IA se hallaba integrada solo por zorros. La disciplina estaba experimentando uno de sus inviernos perennes y la mayoría de los mitólogos se habían escondido. Ray Kurzweil continuaba promocionando su visión cargada de confianza y los teóricos de la IA clásica, como Doug Lenat, seguían dedicados a sus teorías favoritas, buscando la piedra Rosetta de la IA. Pero los altibajos en apariencia interminables habían desgastado buena parte de la disciplina, hasta el extremo de que muchos se sentían incómodos ante la idea de usar siquiera la etiqueta de IA en nuestras investigaciones. La IA se había convertido en una mala publicidad. (Quizá parezca extraño hoy en día, pero es cierto.) El debate no tardó en apuntar hacia el esoterismo de los algoritmos específicos, como las «máquinas de vectores de soporte» y la «máxima entropía», ambos enfoques de aprendizaje automático. Los científicos de la IA clásica los habían descartado como «superficiales» o «empíricos», porque los enfoques estadísticos que usaban datos no utilizaban el conocimiento y no se les daban demasiado bien ni el razonamiento ni la planificación (si es que no se les daban directamente mal). Pero, cuando la red comenzó a proporcionar esos datos tan necesarios, los enfoques comenzaron a revelarse prometedores.

La «revolución» del aprendizaje profundo se inició hacia 2006, con los primeros trabajos de Geoff Hinton, Yann LeCun y Yoshua Bengio. En 2010, Google, Microsoft y otras compañías big tech ya usaban redes neuronales en aplicaciones de consumo masivo como la del reconocimiento de voz, y en 2012 los

smartphones de Android contenían la tecnología de las redes neuronales. A partir de ese momento y hasta 2020 (cuando escribo esto), el aprendizaje profundo ha sido el martillo de todos los clavos en lo que a los problemas de la IA se refiere: problemas que cabría abordar «desde cero», como los juegos y el reconocimiento de voz y los datos de imagen, ocupan la mayor parte de la investigación y de los dólares comerciales en IA.

Con el despegue del aprendizaje profundo, la IA (y la conversación en torno a la IA) tomó vuelo también. Los erizos regresaron y, de manera predecible, los medios atizaron las llamas de aquel nuevo futurismo. Pero algo extraño ha estado pasando últimamente en la IA. Reparé en ello en 2018, a través de voces más escépticas, y en 2019 resultaba ya inconfundible. Están volviendo los zorros.

Muchos mitólogos (con algunas excepciones, pocas, pero notables) no son expertos, como Elon Musk o el fallecido astrofísico Stephen Hawking o incluso Bill Gates. Aun así, ayudaron a generar buena parte del bombo de los medios sobre la IA —principalmente, el bombo del aprendizaje profundo—, que alcanzó su cenit hace poco (alrededor de 2015, año arriba, año abajo). Ahora, no obstante, vuelve a ser cada vez más habitual que se hable de limitaciones —es lo que hace, por ejemplo, Gary Marcus, científico cognitivo y fundador de la compañía de robótica Robust.AI, coautor en 2019, junto con el científico informático Ernest Davis, de Rebooting AI: Building Artificial Intelligence We Can Trust[9] [«El reinicio de la IA: La construcción de una inteligencia artificial en la que podamos confiar»]. Marcus y Davis presentan con argumentos de peso la idea de que la disciplina vuelve a estar sobredimensionada, y de que el aprendizaje profundo tiene sus límites; que se necesitará algún avance fundamental para alcanzar una IA de inteligencia general. En 2017, el científico de IA Hector Levesque (colega de Davis, sobre quien hablaremos con mayor detenimiento más adelante) firmó una provechosa polémica sobre la IA moderna que tituló Common Sense, the Turing Test, and the Quest for the Real AI[10] [«Sentido común, el test de Turing y la búsqueda de la IA real»]. Cuando publiqué el artículo «Questioning the Hype about Artificial Intelligence» [«El furor de la inteligencia artificial, bajo la lupa»] en The Atlantic, en 2015, las reacciones fueron en general desdeñosas.[11] En la actualidad hay más voces críticas, y entre ellas se encuentran las de numerosos líderes reconocidos de la IA, que en efecto cuestionan ese bombo y platillo.

Sigue siendo raro oír argumentos reflexivos sobre la imposibilidad de una IA

real, y eso se debe al mismo motivo por el que la gente se abstiene de realizar predicciones al respecto: porque se desconoce el futuro de la IA. Pero, en lo cultural y en lo psicológico, la disciplina parece haber entrado en una fase de relajación, en la que se advierte a los principiantes y al público expectante de que habrá que recorrer un largo camino para llegar a la inteligencia general. Esta tendencia tiene una importancia capital, porque el mito es un faro emocional mediante el cual navegamos por el tema. Es expansionista, acoge a todas las visitas: conceptos como la consciencia, emociones como la agresividad o el amor, instintos como el sexo y otros ingredientes propios de la mente y de los seres vivos. Pero el nuevo debate «científico» es, más o menos, una narrativa sobre las posibles extensiones de la IA débil hacia una generalización cada vez mayor, de cuyo alcance quedan fuera ideas a gran escala como la de la consciencia. Se pasa de lista, quizás... el mito es lo que motiva el interés de todo el mundo. De otro modo, la cosa se queda en formas de tecnología cada vez más potentes por todas partes, una tendencia que, como enseguida percibimos, tiene dos caras.

HACIA UNA SUPERINTELIGENCIA SIMPLIFICADA

Los errores de inteligencia que ayudaron a forjar nuestro mundo computacional simplificado también han regresado bajo una apariencia moderna. Stuart Russell, coautor de la guía introductoria a la IA definitiva junto con Peter Norvig, de Google, asegura en Human Compatible: Artificial Intelligence and the Problem of Control [«Compatible con la humanidad: La inteligencia artificial y el problema del control»], de 2019, que la inteligencia no implica nada más que alcanzar objetivos —nos ofrece una definición que incluye no solo a los seres humanos, sino a los delfines y las hormigas, a la bacteria E. coli y a los ordenadores—.[12] Además, quiere que se jubile el test de Turing porque ha pasado a ser irrelevante. (Al parecer, mantener una conversación normal y corriente no es un objetivo digno.)

El test de Turing no resulta útil para la IA —escribe— porque se trata de una definición informal y extremadamente accidental, que depende de la

complicadísima y muy desconocida peculiaridad de la mente humana, resultado tanto de la biología como de la cultura. No existe la posibilidad de «vaciar» la definición y comenzar desde cero a crear máquinas que superen el test de manera comprobable. En su lugar, la IA se ha concentrado en el comportamiento racional, [y por tanto] la máquina será inteligente siempre y cuando sus acciones tengan como resultado probable la consecución de lo que quiere, según lo que haya percibido.[13]

Cuesta discutirle a Russell su definición de inteligencia, que lo abarca todo: desde el momento en que Einstein «alcanzó» su «objetivo» al reimaginar la física como algo relativo, hasta la margarita que vuelve su rostro hacia el sol. Pero el rechazo de Russell hacia el test de Turing parece restrictivo y legalista en exceso, ya que el espíritu del test consiste tan solo en que las máquinas que puedan comprender y utilizar un lenguaje natural deben ser inteligentes. En términos prácticos, no deberíamos esperar gran cosa de Siri y demás asistentes personales activados por la voz si nunca se van a enterar de lo que nos están diciendo, así que desdeñar el test no parece muy inteligente. (Si una Siri de próxima generación alcanza algún día el punto de mantener conversaciones ilimitadas y comunes y corrientes con su dueño humano, el test de Turing regresará en cuanto gran «sueño de la IA» realizado al fin. Pero ay...)

Russell, un reconocido experto en IA y profesor de ciencias de la informática en la universidad de California, Berkeley, también se quita de encima el problema de la consciencia: «En el terreno de la consciencia, la verdad es que lo ignoramos todo, así que no pienso decir nada». A continuación nos asegura que «dentro de la IA no hay nadie trabajando en crear máquinas conscientes, ni sabría nadie por dónde comenzar, y ningún comportamiento tiene la consciencia como prerrequisito». Pero, de todos modos, sí dice algo —bastante— acerca de la consciencia:

Supón que te doy un programa y te pregunto: «¿Representa algún tipo de riesgo para la humanidad?». Tú analizas su código y, en efecto, al ejecutarlo, este traza y lleva a cabo un plan cuyo resultado será la destrucción de la raza humana, tal y como un programa de ajedrez traza y lleva a cabo un plan cuyo resultado será la derrota de cualquier ser humano que se enfrente a él. Ahora supón que te digo

que ese código, al ejecutarlo, también genera una forma de consciencia automática. ¿Haría cambiar eso tu predicción? En absoluto. No representa la menor diferencia. Tu predicción acerca de su comportamiento sigue siendo la misma, porque se basa en el código. La verdad es que todos esos argumentos hollywoodienses sobre máquinas que cobran consciencia de manera misteriosa y comienzan a odiar a los seres humanos no han entendido nada: lo que importa es la aptitud, no la consciencia.[14]

Pero quizá sea Russell el que no ha entendido nada, porque la mitología de las máquinas que «cobran vida» conforma en realidad el alma de los sueños acerca de una futura IA. Si informáramos a la gente que llega a la sala de cine para ver un pase de Ex Machina de que el verdadero sueño de la IA consiste en crear superordenadores mecánicos, «sin luces en su interior», que nos ayuden (a nosotros y a nuestros enemigos) a alcanzar objetivos, quizá se quedarían muy poco impresionados. Russell parece sugerir que un sistema algorítmico, animado de la manera adecuada por unos módulos aún desconocidos para obtener una inteligencia general, representará el éxito definitivo de la IA. Los verdaderos erizos entienden aquello que a los zorros cautelosos se les escapa: que la IA incluye la ciencia y el mito, y que su fascinación perdurable en la mentalidad popular implica que se trata de una piedra de toque psicológica y cultural.

Ray Kurzweil ha sostenido en todo momento que, sea lo que sea la consciencia, las máquinas dispondrán de ella a patadas —y que será más rica y «mejor» que la nuestra—. En 1999, su peán al mito llevó el apropiado título de La era de las máquinas espirituales (y lo de «espirituales» lo dijo en serio, como si los ordenadores superinteligentes fueran a tener experiencias conscientes e incorpóreas). Kurzweil insiste sabiamente en que el test de Turing representa una cota de referencia adecuada para la IA: «Para superar el test has de ser inteligente». Incluso le desagrada el concepto de IAG, que cada vez se usa más para hablar de manera específica sobre la inteligencia artificial general, ya que, tal y como afirma (correctamente), «el objetivo de la IA siempre ha consistido en alcanzar grados de inteligencia cada vez mayores, y en última instancia llegar a los niveles humanos de inteligencia».[15]

El del deseo sexual podría ser un tema apropiado para la IA, en cuanto prueba con tornasol de la inteligencia —sobre todo si se trata de un elemento básico de nuestros esfuerzos y anhelos, y para alcanzar objetivos diversos, como parece ser

en la vida—. Ex Machina resulta prácticamente shakespeariana en su combinación de tensión sexual, consciencia, explotación y liberación —y todo ello en una especie de test de Turing—. El novelista reconvertido en realizador Alex Garland aborda la singularidad mientras nos ofrece la historia de una androide superinteligente que planea su huida de la situación de esclavitud a la que la mantiene sometida un científico loco (otro tema cargado de contenido): su solitario inventor, Nathan (interpretado por Oscar Isaac). Presumimos que el objetivo de Ava es superar un test de Turing interactuando con un invitado de Nathan, Caleb Smith (Domhnall Gleeson), y resultando completamente convincente como «humana», pese a que él ha sido informado de que Ava es un androide y lo ve, además—. Se trata del test definitivo, dice Nathan.

Pero Ava tiene sus propias ideas (a la mierda nuestros objetivos) y planea su huida al mundo salvaje, más allá de los confines del centro de investigación de Nathan. Cuando Ava huye al fin, dos humanos están (o estarán) muertos. Al salir ve los colores, unos colores gloriosos, una demostración para el espectador de que está «viva» y consciente de veras.[16] La hemos visto comprender y utilizar el inglés de manera tan efectiva que ha reducido a los dos hombres a un estado de confusión desesperada y derrota. Aquí tenemos, pues, una representación a voz en grito del mito, que muestra la idea futurista y fundamental de un próximo cruce de caminos en el que las máquinas adelantarán a la humanidad, y punto. Ava es más lista, más astuta, más espiritual, y está más viva que sus equivalentes humanos.

La visión de Garland es pura mitología —también es una gran historia humana, que captura temas arquetípicos (la liberación, el bien y el mal, y la sexualidad) a través de la lente de una tecnología futura—. Se da la ironía de que el éxito de Ex Machina se debe a que accede a las más profundas emociones humanas, igual que otras obras de arte como 2001: Una odisea del espacio.

También resulta irónico que algunas voces hayan optado en tiempos recientes por alejar el campo de la IA de su mito feliz (o aterrador) en pos de un debate más «científico» —en otras palabras, han buscado descartar conceptos tan cargados emocionalmente como los de singularidad, consciencia e inteligencia sin dejar de beneficiarse de su presencia en la opinión pública—. Por ejemplo, es evidente que Russell desea separar la labor de corte serio en la inteligencia artificial general de los retratos que hace de ella la cultura popular en películas como Ex Machina. Russell considera que la consciencia es un problema filosófico ridículo (al fin y al cabo, lo ignoramos todo sobre ella), que los tests

de Turing son ideas anticuadas y demasiado vulnerables al engaño, y que cualquier preocupación sobre la posibilidad de que las máquinas se pongan agresivas (o valientes, o emocionales de cualquier otra manera) resulta fundamentalmente errónea. Los ordenadores superinteligentes se limitarán a perseguir sus objetivos. El problema es —y equivale a un riesgo existencial, incluso sin la imaginería propia de Terminator— que sus objetivos podrían no ser los mismos que los nuestros.

Russell admite que ese es un problema que ya tenemos con la IA. En concreto, llama la atención sobre los algoritmos de contenido-selección en la red, cuyo objetivo consiste en maximizar los ingresos publicitarios bombardeando a todo el mundo con anuncios pegajosos y relevantes. Es posible que a la IA superinteligente se le dé demasiado bien perseguir nuestros objetivos. Una metáfora apropiada es la del rey Midas, que obtuvo el poder de convertir todo lo que tocara en oro, y a quien de repente le resultó demasiado fácil convertir todo lo que tocaba en oro, incluyendo a su propia hija (ese no era el objetivo); de manera similar, la superinteligencia a la que se le encomendara un objetivo podría dar con una manera de alcanzarlo que acabara por eliminarnos, quizá incluso sirviéndose de los átomos de carbono de los propios seres humanos como un recurso más, como un medio más para alcanzar su fin.

Esa idea es una preocupación recurrente entre los aprensivos de la superinteligencia. Nick Bostrom plantea un escenario en el que la superinteligencia reciba la tarea en apariencia mundana de maximizar la producción de clips sujetapapeles (el objetivo que le han encomendado los seres humanos), y que poco a poco va convirtiendo todos los elementos del universo en una fábrica de clips sujetapapeles, incluyendo los elementos útiles de nuestro cuerpo. Eliezer Yudkowsky, exdirector del Instituto de Investigación de la Inteligencia de las Máquinas, afirmó una vez en broma: «La IA no te odia, tampoco te ama, pero es que estás hecho de unos átomos que ella podría usar para hacer otra cosa».[17]

La idea de que la superinteligencia venidera estará completamente centrada y será supercompetente en la consecución de un objetivo, pero a la vez mostrará un nulo sentido común, parece ir a contracorriente de lo que representa la superinteligencia misma —de la que se supone, al fin y al cabo, que es la suma de la inteligencia humana con algo más—. Científicos de IA como Gary Marcus, que entienden la inteligencia como algo dotado de (un mínimo de) sentido común (y quienes quizá también dispongan de él), han señalado que a un

ordenador superinteligente que optimizara la fabricación de un producto humano de cara a su venta, como los clips sujetapapeles, también se le podría ocurrir que no debe destruir a los seres humanos, que son sus compradores. De nuevo encontramos una curiosa simplificación de la superinteligencia implícita en las ansiedades de Russell y otros sobre la posibilidad de que esta actúe con un apego ciego e informático hacia el objetivo que le hayan encomendado sus programadores. Es extraño ponerse a vigilar algo así. El propio Russell admite que el sentido común y el lenguaje son hitos capitales que la IA no ha alcanzado aún. ¿A qué se debe su ausencia en esta imagen de la superinteligencia? Que los ordenadores del futuro dispusieran de sentido común permitiría obviar esas preocupaciones, a menos que después de todo acabaran revelándose como agresivos y diabólicos, cosa que a Russell le cuesta descartar como parte de este mito ridículo.

En todo caso, los escenarios del apocalipsis de los clips sujetapapeles sí que preocupan a algunos investigadores de mentalidad científica como Russell, quien sugiere que prevengamos esas posibilidades insertando en los ordenadores superinteligentes del futuro ciertos principios; el primero, asegurarse de que «no adjudican un valor intrínseco» a su propio bienestar, de modo que su único objetivo sea el de maximizar nuestras preferencias. El problema, tal y como nos recuerda Russell, es que a menudo no tenemos ni idea de cuáles son esas preferencias. En última instancia, somos propensos a no formular bien lo que deseamos, en el sentido de lo que le pasó al rey Midas.

Así, junto con el altruismo hacia la humanidad, la IA debe estar impregnada por el principio de la humildad, para desbaratar cualquier error que pueda cometer mientras intenta mostrarse altruista con nosotros (como el de convertir al director ejecutivo de la fábrica de clips sujetapapeles en un clip sujetapapeles llevada por la idea de que está exprimiendo de verdad, pero de verdad, hasta la última gota de su productividad al usar todos los medios que tiene a su alcance). Las máquinas altruistas y humildes ayudarán a protegernos del peligro de que los ejecutivos de empresas tecnológicas fumadores de puros (aunque lo más probable es que ya no los fumen) puedan conferirles motivaciones corruptas, y también de la posibilidad de que las máquinas se vuelvan demasiado listas de manera equivocada, y hagan algo parecido a transformar a todo el mundo en oro. Para mayor confusión, las máquinas «altruistas y humildes» también se acercan mucho al enfoque de Ex Machina sobre la IA: como algo «vivo» después de todo, con una inteligencia real (no solo dedicada a maximizar clips sujetapapeles) y una sensibilidad ética—. Se nos podría perdonar que saquemos

la conclusión de que el debate sobre la IA está condenado a mezclar la ciencia y el mito a perpetuidad.

Hay un tercer principio que Russell considera necesario para abortar la crisis existencial a la que nos conducirá la superinteligencia del futuro: la IA debería desarrollarse de manera tal que aprenda a predecir las preferencias de los seres humanos. En efecto, las máquinas deberían observarnos para aprender más acerca de lo que deseamos, lo que les permitirá descartar algunas acciones que podrían mandarlo todo al demonio, por así decirlo. Conocer las preferencias humanas permitirá que los ordenadores eviten hacernos daño mientras se esfuerzan por conseguir sus objetivos. (Russell no explica por qué deberíamos confiar en una superinteligencia que siga siendo lo bastante idiota para acabar con todos nosotros bajo la impresión de que nos está ayudando y otorgarle el papel de «panóptico» benevolente que no deje de observar y aprender nuestras preferencias.)

La versión de Russell de este relato sobre el riesgo existencial de la IA futurista nos trae a la cabeza los robots universales del dramaturgo checo Karel Čapek, que estaban diseñados para obtener una eficacia óptima en su trabajo y a los que se había despojado de manera deliberada de otros rasgos mentales como el sentido moral, la capacidad de apreciar la belleza y de experimentar sentimientos, y la consciencia. Por algún motivo, los autómatas supuestamente irreflexivos de su drama R. U. R. se sienten contrariados de todos modos y desencadenan una revolución robot que prácticamente acaba con la totalidad de la raza humana. Sin duda, el final de Čapek, es el motivo por el que recordamos esa obra de 1920. Nadie se emociona ante la perspectiva de que una Roomba tuneada aprenda de manera superinteligente (pero irreflexiva —ignora tú esa contradicción—) la mejor manera de aspirar la suciedad, o de limpiar la cocina, o de arreglar el coche. Sin duda resultaría de lo más útil, pero no es a eso a lo que nos referimos como superinteligencia. La que nos emociona es Ava. Una superinteligencia que no tenga consciencia ni sentimientos y que no sea capaz de cometer agresiones diabólicas no tiene nada de inteligente. Si carece de sentido común también será una pobre candidata para nuestras imaginaciones mitológicas. Será una calculadora.

Al unir las inteligencias humana y automática en lo que esencialmente es una búsqueda propia de la teoría de juegos para optimizar objetivos, Russell deja sitio para una visión en apariencia «científica» de la mente informática, pero solo desde la limitación severa de las posibilidades de nuestra propia mente. Este

vuelve a ser un error de inteligencia. La inteligencia humana es diversa, está llena aún de misterios profundos y, hasta donde sabemos, resulta eficazmente ilimitada. Al demoler la inteligencia humana, ligándola a una definición más sumisa ante la informática, el pensamiento contemporáneo sobre la IA tira por la borda una comprensión más rica de la mente. Y nos quedamos con un mundo simplificado.

Quizá ese mundo haga que la conversación sobre la llegada de la inteligencia artificial general parezca más razonable (porque «IAG» no equivale a gran cosa), pero a la vez pone en peligro el interés que existe en el proyecto mismo. Para el caso podríamos retirar la idea completa de superinteligencia y acometer un debate más sincero sobre la posibilidad muy real de que un virus informático con capacidad de destrucción mundial, pongamos que lanzado con evidente mala intención por un grupo de programadores, hiciera caer de manera irreflexiva los mercados financieros o accediera ilegalmente a unos datos capitales para la intimidad de los individuos o la seguridad de los países, y procediera a eliminar tales datos. Esa es una informática que se vuelve efectiva con el descubrimiento de una vulnerabilidad. Es el mundo real, no un mito.

EN RESUMEN

Podemos resumir estas posturas sobre la IA y la gente de la siguiente manera. Los kurzweilianos (mitólogos de la IA, punto) se entusiasman con la idea de que las máquinas posteriores a la singularidad vayan a disponer de conciencia, emociones, motivos y una vasta inteligencia. Con gran ironía, mantienen viva la exploración filosófica trasladando temas shakespearianos a la informática. (Los ordenadores vivirán ricas experiencias espirituales y serán grandes amantes, etc.) Podríamos denominar todo esto como el «efecto Ex Machina».

Los russellianos quieren que Ex Machina no sea más que una película, y reducir la conversación sobre la superinteligencia a ideas más respetables en lo matemático sobre una informática general que alcance «objetivos». Por desgracia, los russellianos también tienden a agrupar a los seres humanos según definiciones restringidas de la inteligencia. Eso reduce la brecha que percibimos entre humanos y máquinas, pero siempre a cambio de limitar las posibilidades de

los primeros. Los russellianos son líderes de pensamiento en una tendencia cultural que he bautizado como «el mundo simplificado». Tal y como dice Jaron Lanier, «una nueva generación ha alcanzado la mayoría de edad con unas expectativas reducidas de lo que puede ser cada persona, y sobre aquello en lo que se puede convertir».[18]

Tanto los kurzweilianos como los russellianos proclaman una visión tecnocéntrica del mundo que a la vez simplifica la visión de la gente —en especial, con sus enfoques deflacionarios sobre la inteligencia como una forma de computación— y expande las visiones de la tecnología al promover un futurismo dedicado a la IA como ciencia y no como mito.

Centrarnos en el batitraje y no en Bruce Wayne nos ha metido en un montón de problemas. Vemos posibilidades ilimitadas para las máquinas, pero un horizonte restringido para nosotros. De hecho, la inteligencia futura de las máquinas es una cuestión científica, no mitológica. Si la IA continúa siguiendo el mismo patrón de superar las expectativas en el mundo falso de los juegos o la colocación de anuncios podríamos acabar, al límite, con unos sabios increíblemente indiscretos y peligrosamente idiotas.

Ahora vamos a centrarnos en la ciencia de la IA, y es aquí —en la investigación científica misma— donde el mundo simplificado se vuelve complejo (y misterioso) de nuevo. Y es que, cuando nos liberamos de las ataduras de nuestros errores de inteligencia, se nos cae la venda de los ojos y un problema verdaderamente formidable se presenta ante nosotros.

Segunda parte

EL PROBLEMA DE LA INFERENCIA

Capítulo 8

No calcules, analiza

La IA es la búsqueda de la inteligencia. A lo largo de los varios capítulos que conforman esta parte del libro espero convencer de que esta búsqueda se enfrenta a obstáculos importantes, que no sabemos cómo superar. Para ello, debemos investigar la naturaleza misma de la inteligencia. Y no hay nada mejor para iniciar nuestra investigación que acompañarnos de un «joven extraño e interesante», el detective aficionado Auguste Dupin, a quien nos presenta un narrador sin nombre en el que quizá sea el primer relato de detectives de la historia, «Los crímenes de la calle Morgue».[1]

SOBRE LA RESOLUCIÓN DE CRÍMENES

El narrador —que comparte numerosos rasgos con Edgar Allan Poe, el autor— nos cuenta desde un principio que está obsesionado con los métodos de pensamiento. Siente curiosidad por la manera en que la mente humana conecta fragmentos de información que en apariencia no tienen relación entre sí a partir de la observación cuidadosa y el razonamiento —a partir de inferencias—. Menuda casualidad, pues, que el narrador se encuentre alojado en una vieja casa con Dupin, y que pueda pasarse todo el día cerca de tan brillante detective.

No tardamos en averiguar que Dupin no es un sujeto normal. Posee ese tipo de personalidad extraña de quienes son verdaderamente originales. Y vaya si es extraño. Dupin procede de una familia distinguida, pero se ha visto rebajado a un estado cercano a la pobreza, cosa a la que apenas presta atención porque está pensando constantemente, perdido en sus ideas. Cuando habla se debe a que está rumiando en voz alta. Por supuesto, eso podría acabar resultando molesto. Pero el narrador siente gran aprecio por su «peculiar capacidad analítica». Dice: «Nos pasábamos el día leyendo, escribiendo o conversando, hasta que el reloj nos

advertía de la llegada de la Oscuridad verdadera. Entonces salíamos a las calles cogidos del brazo, seguíamos hablando sobre los temas del día o vagábamos de aquí para allá hasta una hora tardía, en busca, entre las luces y sombras salvajes de la ciudad atestada, de la excitación mental sin fin que la observación callada suele procurar».

Dupin es un prototipo, un modelo, como Sherlock Holmes. E, igual que Holmes, repara en lo que la policía se las arregla para pasar por alto en la aplicación de su «simple diligencia y actividad».

Una noche, a solas en la vieja casa con Dupin, el narrador coge la edición vespertina de la Gazette des tribunaux y se entera de los asesinatos de la calle Morgue:

Crímenes extraordinarios. — La madrugada pasada, alrededor de las tres, los vecinos del Quartier St. Roch despertaron de su sueño a causa de una sucesión de alaridos terribles, procedentes, al parecer, del cuarto piso de una casa de la calle Morgue de la que solo constaban como ocupantes una tal Madame L'Espanaye y su hija, Mademoiselle Camille L'Espanaye. Con cierto retraso, debido a un intento infructuoso por entrar de la manera tradicional, se utilizó una palanca para romper la puerta y entre ocho y diez vecinos accedieron al lugar acompañados de dos gendarmes. En aquel momento los gritos ya habían cesado pero, mientras el grupo subía a la carrera el primer tramo de escaleras, procedentes al parecer de la parte superior del edificio, se pudieron distinguir dos o más voces roncas, que mantenían una discusión furiosa. Al llegar al segundo descansillo, los sonidos también habían cesado y reinaba un silencio perfecto. El grupo se dispersó, sus miembros se apresuraron a entrar en una habitación tras otra. Al llegar a la espaciosa sala trasera del cuarto piso (la puerta de la cual hubo que forzar, al estar cerrada con llave y encontrarse la llave del lado de dentro), el espectáculo que se desplegó ante los presentes sumió a cada uno de ellos no tanto en el horror como en la perplejidad.

El apartamento se encontraba completamente revuelto... los muebles estaban rotos, los habían arrojado en todas direcciones. Solo había una armadura de cama, a la que le habían quitado el colchón para arrojarlo en medio del suelo. Sobre una silla descansaba una navaja, manchada de sangre. En la chimenea había dos o tres mechones largos, gruesos y canosos de cabello humano, también

salpicados de sangre y que parecían haber sido arrancados de raíz. En el suelo se encontraron cuatro Napoleones, un pendiente de topacio, tres cucharas grandes de plata, otras tres más pequeñas de argentán y dos bolsas que contenían cerca de cuatro mil francos en oro. Los cajones del secreter, en una esquina, estaban abiertos y al parecer los habían registrado, aunque en ellos quedaban numerosos artículos. Debajo de la cama (no debajo del armazón de la cama) encontraron una pequeña caja fuerte de hierro. Estaba abierta, la llave aún en la puerta, y no contenía nada más que algunas cartas viejas y otros documentos de escasa importancia.

No había rastro de Madame L'Espanaye pero, al reparar en la inusual cantidad de hollín que se amontonaba en el hogar, se realizó una búsqueda en la chimenea y (¡cuán horrible es contar esto!) de su interior sacaron a rastras el cadáver de la hija, que colgaba boca abajo. Lo habían introducido a la fuerza por la estrecha abertura y había recorrido una distancia considerable. El cuerpo aún conservaba buena parte de su calor. Tras examinarlo se reparó en que había sufrido diversas escoriaciones, sin duda debidas a la violencia con que la habían metido chimenea arriba y después retirado. Tenía varios rasguños fuertes en la cara y unos cardenales oscuros en la garganta, y las marcas profundas de unas uñas, como si la fallecida hubiera sido asfixiada hasta la muerte...

Tras una investigación minuciosa de cada parte de la casa, al no descubrirse nada más, el grupo se dirigió hacia un pequeño jardín empedrado en la parte trasera del edificio, donde yacía el cadáver de la anciana, que había sufrido un corte tan profundo en la garganta que, en el intento por levantarla, la cabeza rodó por el suelo. El cuerpo, igual que la cabeza, había sido terriblemente mutilado —el primero, a un extremo tal que a duras penas conservaba apariencia alguna de humanidad.

De momento, entendemos que no existe la más ligera pista sobre este misterio horrible.[2]

Al día siguiente, la Gazette publica nuevos detalles sobre el caso. A partir de los recuentos de quienes prestan declaración podemos reunir la información relevante. Madre e hija eran pudientes. Tres días antes de los asesinatos, la madre había retirado una suma importante de dinero del banco, en oro, que tras los crímenes se encontró a plena vista, sin tocar, en el suelo. También resulta

curioso que uno de los policías del primer grupo que llega a la escena manifieste haber oído dos voces; una, de manera evidente, de un hombre que hablaba francés, y otra que no pudo reconocer en absoluto y que definió como «severa, aguda y muy extraña». Pensó que se trataría de un extranjero, posiblemente español. Más adelante, otros testigos describirían la voz como posiblemente italiana, rusa o inglesa.

Era desconcertante. El dinero —quizá el motivo más probable para cometer un asesinato— se queda sin tocar en la casa. Las puertas están cerradas con llave por dentro. El cuerpo de la hija se encuentra en el interior de la chimenea, alojado en ella con tanta fuerza que es necesaria más de una persona para sacarlo de allí. Y las voces también. Las de los asesinos, al parecer, pero, aunque la policía oyó con claridad, mientras subían las escaleras de la casa, que eran dos, solo una resulta reconocible, y afirman que la otra es una mezcla extraña de jerigonza. Ninguno de los testigos es capaz de decir con exactitud lo que se dijo (si se dijo algo), ni en qué idioma.

La policía está perpleja. Las declaraciones de los testigos no hacen más que ahondar su confusión. Con ello vengo a decir que todas las pistas, en su conjunto, no apuntan hacia ninguna parte. Los crímenes son un misterio, y ese es el motivo por el que nuestro excéntrico detective aficionado, Dupin, siente un interés tan entusiasta.

El narrador sugiere que Dupin resuelve el caso muy pronto, tras leer en el periódico el comunicado de la policía. No obstante, los dos reciben permiso para visitar la vieja casa de la calle Morgue con la escena del crimen aún intacta. De regreso a casa, Dupin se detiene en las oficinas de otro periódico y pone un anuncio en la sección de objetos perdidos. ¿Alguien en París, presumiblemente el marinero de un navío maltés, ha perdido a su orangután? Existe la posibilidad de que el dueño llame para reclamarlo.

Y aquí aparece la inferencia que permite resolver el caso de los crímenes de la calle Morgue: ningún ser humano mató a la anciana y a su hija aquella noche. El asesino no fue una persona, sino un animal salvaje al que un marinero se trajo de la jungla y mantuvo en alguna morada cercana. Frenético, habiendo escapado de su amo, el orangután, tras balancearse en el postigo externo, entró de un salto por la ventana de la casa profiriendo chillidos y alaridos y esgrimiendo una navaja de afeitar. Ahí está el arma del crimen: en la navaja que decapita a la anciana y en la pura fuerza bruta del animal, que mete a la hija en la chimenea

arrastrándola por los pies.

¿Y la voz humana que oyeron los testigos? El dueño del orangután. ¿Y los sonidos amortiguados e ininteligibles? Los gruñidos del animal.

EL MÉTODO CONJETURAL

Pero ¿quién inferiría eso a partir de los datos del caso? Sin duda, todo el mundo los tiene delante de los ojos. En realidad, Dupin se limitó a conjeturar. La policía siguió los métodos tradicionales hasta que estos no desembocaron en nada. A partir de ese momento también comenzaron a hacer conjeturas. La única diferencia radicó en que la de Dupin fue la buena.

Poe comienza «Los crímenes de la calle Morgue» reflexionando sobre la naturaleza del pensamiento. La historia ficticia del crimen se inicia desde la no ficción. El autor busca las palabras adecuadas. El razonamiento de Dupin, decide, es un triunfo del análisis, en contraste con cálculos más formalistas. El cálculo consiste en conectar unos puntos que ya se conocen; en aplicar las reglas del álgebra, pongamos. El análisis consiste en ofrecer sentido a esos puntos, dando saltos o efectuando suposiciones que los expliquen —y, a continuación, desde una cierta perspectiva, usando el cálculo para comprobarlas—. El cálculo tiene sus límites: «Pero es en aquellas cuestiones que trascienden los límites de la mera regla cuando se demuestra la habilidad del analista». Seguir las reglas no es suficiente, pero no queda claro qué más hace falta con exactitud. El aprecio de Poe hacia ese misterio se vuelve evidente en la declaración que realiza al comienzo de la historia: «Las condiciones mentales que suelen considerarse analíticas son, en sí mismas, poco susceptibles al análisis».[3]

Algunas décadas más tarde, el científico y filósofo norteamericano Charles Sanders Peirce iba a leer los relatos de Poe con gran fascinación. Peirce también se preguntó por la manera en que pensamos, en que razonamos acerca de las cosas. Incluso se las arregló para capturar la gimnasia mental de Dupin en forma de símbolos lógicos. No sabía cómo automatizar el perspicaz estilo conjetural del detective, pero pensó que se trataba de un aspecto básico del pensamiento humano en general.

Para Peirce, el pensamiento no es un cálculo sino un salto, una suposición. No hay ninguna certeza. Nos dedicamos a juntar las piezas. Explicamos y revisamos. Al haber vivido durante el siglo XIX, Peirce no llegó a conocer los ordenadores digitales, pero anticipó lo que iba a hacer de la IA un problema tan complicado para todo el mundo. En realidad, se reduce a esto: puesto que nuestro propio pensamiento se basa en una serie desconcertante de conjeturas, ¿qué esperanza hay de que podamos programarlas?

Con el tiempo, Peirce acabaría desarrollando todo un marco explicativo para el razonamiento humano que se basó en la lógica formal y sus clases, como la deducción y la inducción.

Pero había un tercer elemento, reflexionó Peirce, que capturaba nuestros juegos conjeturales. Lo llamó «abducción», y en él nos vamos a centrar a continuación.

Capítulo 9

El puzle de Peirce (y el rompecabezas de Peirce)

Para aquellos que estén familiarizados con su trabajo, Charles Sanders Peirce forma parte de un selecto grupo de pensadores importantes y verdaderamente originales: En la biografía Charles Sanders Peirce: A Life [«Charles Sanders Peirce. Una vida»], el historiador Joseph Brent se refiere a él como, «quizá, la mente más importante que Estados Unidos haya producido nunca». En 1934, el filósofo Paul Weiss describe a Peirce en The Dictionary of American Biography [«Diccionario de biografías norteamericanas»] como «el más original y versátil de los filósofos norteamericanos, y el más importante lógico del país». El crítico e historiador cultural Lewis Mumford le situó en compañía de genios iconoclastas como Roger Bacon y Leonardo da Vinci. Y, durante una entrevista de 1976, cuando le preguntaron al pionero de la ciencia lingüística del MIT Noam Chomsky por sus influencias, este contestó: «En relación con las cuestiones que hemos estado tratando [relativas a la filosofía del lenguaje], el filósofo que siento más cercano y aquel al que prácticamente parafraseo es Charles Sanders Peirce».[1]

BRILLANTE PERO ABANDONADO

Como Albert Einstein, Peirce era zurdo y pensaba en imágenes. Bosquejaba inferencias lógicas en forma de diagramas. Pasó los últimos años de su vida escribiendo solo en casa, quejándose de que pasaba hambre y frío, y de que era demasiado pobre para permitirse el combustible del fogón. Los pocos amigos que tenía se preocupaban por él y se las apañaron para conseguirle una serie de conferencias en Harvard sobre los fundamentos de la lógica, en las que Peirce explicó los tipos de inferencia lógica desde un marco que, según él, afianzaba el método científico —un método para pensar con claridad—. Entre los asistentes se encontraba William James, el famoso filósofo y precursor de la psicología de

Harvard, quien más tarde confesó que no había acabado de entender las conferencias, que las matemáticas relacionadas con las imágenes y diagramas de Peirce se encontraban más allá de su comprensión. Al parecer, James no fue el único a quien le sucedió algo así; las conferencias pasaron mayormente desapercibidas y solo aparecieron en forma de libro décadas después.

Nacido en el entorno de la cultura científica victoriana de Cambridge, Massachusetts, en 1839, Peirce fue miembro de una familia pudiente y destacada en diferentes ámbitos. Su padre fue un eminente profesor de matemáticas en Harvard. Un primo más joven, Henry Cabot Lodge, se convirtió en un poderoso senador. Peirce recibió una educación clásica y en 1863 se graduó summa cum laude en la Escuela Científica Lawrence de la universidad de Harvard. Pasó treinta años trabajando como investigador científico en el Servicio de Costas y en el Servicio Geodésico Nacional, donde realizaba estudios topológicos de la superficie terrestre usando precisas medidas gravimétricas. Fue químico aficionado, un prestigioso conferenciante sobre lógica en la universidad Johns Hopkins y el primer delegado norteamericano de una asociación científica internacional. Fue científico, lógico, filósofo, escritor, prolífico reseñista de libros para el Nation, y más. Max H. Fisch, estudioso de Peirce que se pasó décadas investigando su vida y obra, ofrece este adecuado juicio grandilocuente sobre sus numerosos logros:

¿A quién pertenece el intelecto más original y versátil que hayan generado hasta el momento las Américas? La respuesta «Charles S. Peirce» no encuentra oposición, porque cualquier nombre que citáramos en segundo puesto se hallaría tan lejos que no valdría la pena ni mencionarlo. Matemático, astrónomo, químico, geodesta, topógrafo, cartógrafo, metrólogo, espectrólogo, ingeniero, inventor; psicólogo, filólogo, lexicógrafo, historiador de la ciencia, economista matemático, estudioso de la medicina a lo largo de toda su vida; crítico de libros, dramaturgo, actor, autor de relatos; fenomenólogo, semiótico, lógico, retórico, metafísico... Es el único filósofo de sistemas de las Américas que se ha mostrado a la vez competente y productivo en lógica, matemáticas y en una amplia gama de ciencias. Si en ese sentido ha tenido algún par a lo largo de toda la historia de la filosofía, no habrán sido más de dos.[2]

Sin embargo, Peirce murió como un paria, en gran medida olvidado. Los genios olvidados han sido tan comunes a lo largo de la historia que a veces los acabamos redescubriendo, como sucedió con Tesla. Pero, sin duda más que Tesla —quien, al fin y al cabo, alcanzó una especie de fama póstuma como motivo de inspiración de Elon Musk para bautizar una compañía de coches eléctricos en su honor—, Peirce sigue siendo un pensador importante al que los libros de historia no suelen mencionar. Su trabajo ha sido apreciado sobre todo en la filosofía, donde se le recuerda como el fundador de la escuela conocida como pragmatismo.

Sus obras tempranas sobre computación están completamente olvidadas. Los académicos continúan sondando sus textos voluminosos sobre la naturaleza de la lógica, pero se trata de un tema oscuro, demasiado difícil para conectarlo con el debate general. Por ello, aunque quienes han comprendido la importancia y el relieve de sus ideas sobre la naturaleza de la lógica le hayan comparado con Aristóteles, el debate contemporáneo sobre las ideas de Peirce requiere en la mayoría de los círculos un bosquejo biográfico... y una explicación, incluso una disculpa.

FÍSICA, FILOSOFÍA Y PERSONALIDAD

¿Por qué se ha olvidado a Peirce? Su vida personal nos ofrece una pista: logró irritar a casi todo el mundo. William James no dejó de ser su amigo íntimo de por vida, pero incluso él salió de su primer encuentro con Peirce, cuando ambos eran alumnos de Harvard, con sensaciones encontradas, tal y como le contó por carta a su familia: «Está el hijo del profesor Peirce, de quien sospecho que es un “tipo” muy inteligente y con mucho carácter, bastante independiente pero también agresivo».[3] Comprensivo, James más tarde se refirió a Peirce como «ese ser extraño y rebelde».[4]

La personalidad espinosa de Peirce y su indiferencia hacia las costumbres de sus contemporáneos le granjearon problemas sin fin, tanto en lo personal como en lo profesional. Solía ofender a los miembros de la alta sociedad victoriana de Nueva Inglaterra (incluyendo a su familia), quienes le rehuían justificadamente. Harvard se negó a ofrecerle una cátedra a causa de una conocida infidelidad en

su matrimonio. El Servicio de Costas de Estados Unidos, para el que trabajó durante décadas, acabó despidiéndole por no entregar sus informes a tiempo y por perder un equipo de gran valor mientras viajaba por Europa. También le despidieron de la Johns Hopkins tras unas denuncias indeterminadas relacionadas con una conducta impropia. Hoy diríamos que no encajaba en los sitios, que fue el estereotipo perfecto del genio incomprendido. Su constitución era incapaz de respetar las reglas del juego.[5]

Los escándalos e idiosincrasias personales de Peirce ayudan a explicar que, por ejemplo, los registros de su vida privada —volúmenes de documentos— permanecieran sellados en la biblioteca Houghton de Harvard hasta 1956, cuarenta y dos años después de su muerte. Sus artículos científicos y filosóficos —incluyendo muchos de un interés enorme para la ciencia computacional y, en especial, la IA— se quedaron relativamente intactos, e inéditos, en los archivos de Harvard, a falta de «unos pocos miles de dólares» como «garantía por los gastos iniciales de su publicación», en palabras de Lewis Mumford.[6] Al no entender sus ideas y su valor, por su deseo de no atraer el escándalo, muchos de quienes conocieron a Peirce jamás fueron testigos del restablecimiento de su reputación, ni de la publicación de buena parte de la obra de su vida. Peirce mismo murió en el olvido; le sobrevivió Juliette, su no menos enigmática segunda esposa, una francesa que también había llevado una vida accidentada. De manera apropiada, la familia de Peirce a veces describía a Juliette como una paria, o una «gitana».

Solo mucho, mucho tiempo después hemos llegado a comprender la enormidad de la contribución de Peirce a las matemáticas y, en especial, a la lógica —hasta cierto punto, ni siquiera hemos acabado de hacerlo—. Son de la máxima importancia sus ideas sobre la inferencia lógica y, en particular, para desplazarnos hacia la atracción principal de su trabajo vital, su exploración de la profundidad y el misterio de lo que él denominó la «inferencia abductiva», una especie de conjetura explicativa que, según se dio cuenta, afianza la mayor parte de nuestro pensamiento.

Peirce reparó en el hecho de que el razonamiento abductivo había quedado fuera de todos los recuentos de razonamiento lógico desde tiempos de Aristóteles. Tampoco encajaba en el marco lógico habitual que asumían las matemáticas y los cursos de lógica. Vio la abducción como una pieza lógica perdida que planteaba preguntas fundamentales sobre la automatización y la inteligencia. Si hubiera sabido de la IA, lo más probable es que hubiera visto algo que todavía

hoy se sigue pasando por alto a menudo: que el problema de la inferencia abductiva confronta a la IA con su desafío central, aún no resuelto del todo.

EL PUZLE DE LA INFERENCIA

El narrador de Edgar Allan Poe se devanó los sesos buscando las palabras que describieran aquello a lo que Peirce más tarde dedicaría diversos volúmenes: la inferencia abductiva. Pero la inferencia abductiva es un tipo de inferencia. ¿Qué es una inferencia? Un nombre, para comenzar. La forma verbal es «inferir», que se refiere a una acción. Etimológicamente, «inferir» significa «traer aparejado», del latín «in», en, y «ferre», traer. El Oxford English Dictionary nos dice que se trata de algo que hacemos cognitivamente, con la mente: «Alcanzar una opinión o decidir que algo es verdadero basándose en la información de la que se dispone».

Por desgracia, el OED también dice que «deducir» es sinónimo de inferir, lo cual no nos ayuda demasiado (porque la deducción es solo una de las formas de la inferencia).

El OED también ofrece algunos ejemplos de uso que recalcan la generalidad de la palabra «inferir» en el lenguaje cotidiano.

Inferir algo (a partir de algo): Hay que inferir buena parte del sentido a partir del contexto. Inferir los motivos del asesino queda en manos de los lectores.

Inferir que: Resulta razonable inferir que el gobierno conocía esos acuerdos.

La inferencia consiste en dar lugar a una nueva idea, lo cual en lógica equivale a extraer una conclusión y, de manera más general, implica usar aquello que ya sabemos, y lo que vemos u observamos, para actualizar convicciones previas. Podríamos inferir los motivos del asesino (por tomar prestado uno de los ejemplos del OED) sirviéndonos de lo que ya sabemos y de lo que hemos leído

en los periódicos (igual que Dupin).

La inferencia también es una especie de salto, pero un salto que estimamos razonable, como cuando inferimos que «el gobierno conocía esos acuerdos» —de nuevo, de acuerdo con los conocimientos previos de los que dispusiéramos (conocimiento público o compartido) junto con la lectura (observación) de alguna historia o historias de actualidad.

La inferencia es un acto cognitivo básico de la mente inteligente. Cuando un agente cognitivo (una persona, un sistema de IA) no es inteligente, sus inferencias serán incorrectas. Pero todo sistema que infiera algo debe disponer de una inteligencia básica, porque el acto mismo de utilizar lo que se sabe y lo que se observa para actualizar nuestras convicciones se encuentra ligado de manera inevitable a lo que queremos transmitir al hablar de inteligencia. El sistema de IA que no infiere nada no merece recibir el nombre de IA. (Aunque podríamos decir que incluso un sistema encargado de etiquetar fotos de gatos infiere que lo que «ve» es un gato, así que el listón puede encontrarse bastante bajo.)

Es imposible pillar un chiste, descubrir una nueva vacuna, resolver un asesinato como hace Dupin o simplemente mantenerse al día de los diversos acontecimientos y comunicaciones que se dan en el mundo sin algún tipo de capacidad de inferencia. Sabemos un montón de cosas, sin duda, pero solo las inferencias nos permiten obtener conocimientos (o convicciones) nuevos. Sabemos que mañana saldrá el sol, así que no necesitamos inferirlo. Asimismo, no nos molestamos en inferir que seguimos teniendo la mano pegada al brazo. Es un conocimiento del que ya disponemos, un conjunto de convicciones que ya hemos formado. Pero nuestro conocimiento cambia de manera constante, se actualiza. Si vemos por la ventana que ha oscurecido misteriosamente, pues es demasiado temprano para ello, podemos inferir que se está produciendo un eclipse solar, o quizá que una enorme tormenta de arena ha ocultado el sol por el oeste, o quizá que ha habido un holocausto nuclear. Todo depende de... ¿qué sabemos en ese momento? Por lo que vemos, ¿qué es lo que tiene más sentido?

En un sentido amplio, siempre estamos infiriendo; es como una de las condiciones de estar despierto. Puedo ir a la cocina, encontrarme con una lata medio vacía de Pepsi e inferir que mi hermana la ha dejado allí, ya que bebe Pepsi y está de visita. Por otro lado, hay unos operarios reformando las encimeras, y antes reparé en que uno de ellos se bebía una Pepsi. Es más, hace

un rato estuve bebiéndome una Pepsi y la dejé sin terminar en el porche, así que es posible que mi esposa la haya entrado a la casa. Acabamos conjeturando una explicación que tenga sentido, dado lo que sabemos y el contexto en el que nos encontramos. Se trata de una inferencia «en tiempo real», ya que extraemos esas conclusiones mientras entramos en la cocina. Las circunstancias del mundo real no dejan de cambiar, así que la inferencia en tiempo real es normal. Al fin y al cabo, pensamos dentro del tiempo. Un programa informático que tarde diez mil millones de años en resolver un problema no será para nada inteligente, y tampoco lo será el que, en tiempo real, se estrelle contra una pared.

La naturaleza provisional de numerosas inferencias implica que las iniciales puedan estar equivocadas, sobre todo cuando nos hemos apresurado a alcanzarlas. Si llego tarde a la oficina, el jefe puede inferir que no me tomo las cosas con seriedad, cuando de hecho me he encontrado con un atasco de tráfico a causa de un accidente. En otras palabras, el jefe ha extraído una conclusión basada en una impresión preformada o prejuicio sobre mí. La gente usa la palabra «inferencia» en sus conversaciones cotidianas otorgándole ese sentido, refiriéndose a un salto apresurado hacia una conclusión injustificada: «Oh, es Suzy, que está infiriendo todo tipo de locuras sobre ti después de lo que dijiste anoche». Y es cierto que, en un sentido técnico, Suzy está realizando inferencias, pero la implicación aquí es que se trata de inferencias tendenciosas, y que Suzy se encuentra demasiado dispuesta a hacer suposiciones injustas (quizá porque está de mal humor o porque no le caes bien).

En un sentido más específico, la inferencia ingresó en el léxico matemático hace mucho tiempo, y en épocas más recientes ha participado en los debates sobre la informática y la IA. En ese contexto, la «inferencia en tiempo real» puede referirse a un robot que se oriente por un entorno dinámico, como una calle bulliciosa. La «inferencia probabilística» extrae conclusiones de datos estadísticos y tiene una aplicación evidente en los enfoques de IA basados en datos.

Hubo un tiempo en que los científicos de IA se pelearon vigorosamente con una condición previa de la inferencia, cuyo uso inteligente ya conocemos: la cuestión del «conocimiento». Los sistemas que no saben nada tampoco pueden inferir gran cosa. Así que aquellos primeros investigadores intentaron programar el conocimiento en los sistemas de IA, para ayudarlos a que dieran sentido a su sensor o a sus impulsos de texto. Se descubrió (por las malas, a través de fracasos repetidos) que los sistemas de IA con amplios depósitos de

conocimiento en forma de datos y reglas seguían necesitando usar ese conocimiento dentro de un contexto para extraer conclusiones relevantes. Ese «uso» del conocimiento es lo que lleva a que las inferencias sean tan difíciles. ¿Qué fragmento de conocimiento resulta relevante en el pajar de mi memoria para aplicarlo al mundo dinámico y cambiante que me rodea?

La capacidad de determinar los fragmentos de conocimiento con relevancia no es una aptitud informática. Poe insiste en que, en el reino de lo «analítico», no existen fórmulas que permitan llegar a la percepción humana; se trata de «asuntos que van más allá de los límites de la mera regla» o cálculo. En efecto, Dupin parece llegar a una explicación para los crímenes —el orangután— a través de una conjetura casual, que más tarde verifica al encontrarse con el dueño del animal perdido. Por tanto, ¿se trató solo de una suposición por su parte? En un sentido importante, sí. Pero eso no anula su carácter de inferencia. Hace que esta cobre importancia.

MÁS TURING

En su artículo seminal de 1950, «Maquinaria computacional e inteligencia», Turing desestimó la cuestión de que las máquinas llegaran a pensar burlándose de su propio título y asegurando que el «pensamiento» es subjetivo y carece por completo de rigor científico. Decir que los ordenadores piensan es como decir que los submarinos nadan. Que se hable de «nadar» ya es un ejercicio de antropomorfismo. Los delfines nadan, pero los submarinos no. Turing pensó que con el uso de la palabra «pensamiento» sucedía algo parecido. Cuando un ordenador juega al ajedrez, ¿quién puede decir si está pensando o si se limita a hacer cálculos?

A Turing le interesaba una mente completamente programable. Por consiguiente, desechó la distinción que había realizado en un principio entre intuición e ingenio llevando la intuición —fuera lo que fuese— a la esfera de la computación. De ese modo, logró que el problema de la IA fuera totalmente comprobable. La tesis era radical incluso según sus propios parámetros previos, pero no se lo tendremos en cuenta porque sentó las bases para que, más tarde, en esa misma década, los investigadores de IA se pusieran a trabajar sin

preocupaciones filosóficas que retrasaran sus avances.

Por desgracia, lo que nunca se abordó debidamente fue la manera exacta en que la inferencia informática podía ser igual que —o podía convertirse en— una inferencia humana. La disciplina no comenzó con una teoría de la inferencia, lo que podría haber proporcionado un plano completo para su desarrollo futuro (o la demostración de su imposibilidad). Que los investigadores de IA carezcan de una teoría de la inferencia es como si los ingenieros nucleares se hubieran puesto a trabajar en la bomba nuclear sin haber resuelto antes los detalles de las reacciones de fisión. Es evidente que no basta con saberse la ecuación de Einstein. Y que tampoco basta con que los entusiastas de la IA tengan conocimientos de teoría de la computación, porque la pregunta misma a la que se enfrentan los científicos que trabajan con la IA es cómo se puede llevar la computación al rango adecuado y a los tipos de inferencia que exhibe la mente. Era algo que se tendría que haber preguntado desde un principio. Al ignorar la pregunta o esquivarla, la disciplina generó falsas esperanzas, condujo a callejones sin salida y perdió el tiempo de manera inevitable.

Y es que hay mucho sobre lo que reflexionar. Tomemos, por ejemplo, las numerosas inferencias que encontramos en la historia de la ciencia. Los científicos bosquejan hipótesis, y a continuación las ponen a prueba. Pero a esas hipótesis no se ha llegado de manera mecánica; son famosas por brotar en la cabeza de los científicos (por lo general, después de que estos hayan alcanzado la maestría en su disciplina). Tal y como hizo Turing una vez, los estudiantes de esos descubrimientos científicos tienden a apartar los saltos intelectuales de las formalidades de la práctica científica, de modo que el acto central de inteligencia «se suma al viaje sin pagar por él» —y se queda sin que lo sometan a análisis—. Pero esas hipótesis son actos genuinos de la mente, primordiales para cualquier ciencia, y a menudo no se los puede explicar señalando los datos o las pruebas o cualquier cosa evidente o programable.[7]

Al plantear que la Tierra giraba alrededor del Sol, y no al revés, Copérnico ignoró montañas de datos y pruebas acumuladas a lo largo de los siglos por los astrónomos que habían trabajado con el antiguo modelo ptolemaico. Lo redibujó todo, con el Sol en el centro, y calculó un modelo heliocéntrico que fuera útil. Aún más importante, el modelo copernicano resultó de hecho menos predictivo pese a ser correcto. En un principio se trató solo de un armazón que, en caso de completarse, podría ofrecer una serie de explicaciones elegantes para reemplazar otras cada vez más enrevesadas, como la aparente retrogradación de los planetas,

que era una pesadilla para el modelo ptolemaico. Solo ignorando en un principio todos los datos, o reconceptualizándolos, pudo Copérnico rechazar el modelo geocéntrico e inferir una estructura radicalmente nueva para el Sistema Solar. (Y fíjate en que eso plantea la pregunta de cómo le habrían podido ayudar los big data, dado que todos los datos encajaban con el modelo equivocado.)

El salto copernicano que hizo despegar la revolución científica podría describirse mejor como una suposición inspirada. Y lo mismo podría decirse acerca de la elección por parte de Kepler de la elipse para describir el movimiento de los planetas, porque en las órbitas planetarias se puede encajar un número muy grande (técnicamente infinito) de figuras geométricas (quizá excluyendo las de tipo trascendental, como las sinusoides). La elipse no era una solución más sencilla que las demás —no se trató de una explicación del tipo navaja de Occam —. Kepler literalmente conjeturó una explicación que a él le «parecía correcta». Que las conjeturas desemboquen en descubrimientos no encaja en el relato mecánico de la ciencia; más bien lo contradice. Pero la labor del detective, los descubrimientos científicos, la innovación y el sentido común son todos obra de la mente; son todos inferencias que los científicos de IA que buscan máquinas de inteligencia general deben explicar de alguna manera.

Como se puede ver, la creación de modelos cognitivos —la construcción de un ordenador que piense, que infiera— resulta desconcertante. Los investigadores en IA (al menos de momento) deberían preocuparse sobre todo de la inferencia en su contexto cotidiano. ¿Por qué? Pues porque la vasta mayoría de las inferencias que realizamos son aparentemente mundanas, como las suposiciones y saltos variopintos que efectuamos en el transcurso de una conversación común y corriente. Por desgracia para los investigadores en IA, ni siquiera esas inferencias mundanas son fáciles de programar. El test de Turing, por ejemplo, resulta complicado en esencia porque la comprensión del lenguaje natural requiere un montón de inferencias de sentido común, que ni son ciertas lógicamente ni resultan (a menudo) demasiado probables. Requiere, en otras palabras, muchísimas abducciones.

Por lo general ni siquiera reparamos en esas inferencias, lo cual está bien: en caso de hacerlo, tenderíamos a quedarnos atascados en bucles solipsistas, dándoles vueltas a las cosas en nuestras cabezas. Esto nos devuelve a Peirce y, de manera más específica, nos conduce al marco tripartito de inferencias en el que se afianza la inteligencia: deducción, inducción y abducción.

Capítulo 10

Problemas de deducción e inducción

A lo largo de la mayor parte de la historia intelectual, la inferencia ha sido sinónimo de deducción. Aristóteles estudió una forma simple de la deducción conocida como el silogismo —dos proposiciones de las que se sabe o se cree que son ciertas conducen a una tercera, la conclusión—. Aristóteles desarrolló una forma temprana de lógica usando silogismos para analizar los argumentos que realizaban él y otros, y para sentar las bases del razonamiento correcto. En su tradición, la inteligencia debe cumplir con las reglas conocidas de la deducción.

Eso tiene sentido. No deberíamos dejarnos persuadir, por ejemplo, por una persona que argumentara que Ray Charles es Dios porque «Dios es amor» y «El amor es ciego» (igual que Ray Charles). Ese argumento es falaz; quebranta las reglas del razonamiento deductivo. Precisamente, anotar todo eso forma parte de la tradición de la lógica deductiva. Aristóteles también exploró la relación entre las reglas deductivas y el llamado razonamiento práctico —por ejemplo, cuando un agente inteligente formula un plan para alcanzar un objetivo cuyos pasos se pueden analizar lógicamente—. (El plan podría ser «correcto» de forma comprobable y, sin embargo, fracasar durante su ejecución; aun así, sería un principio.)

El razonamiento lógico (correcto) y la planificación son subcampos importantes de la IA, y la IA clásica exploró, casi desde sus inicios, estrategias de razonamiento y planificación que usaban elementos de la lógica simbólica, como la deducción. Por ejemplo, un sistema de IA puede implementar un silogismo y también planificar un algoritmo (reglas de la forma: {A, B, C,...} → G, donde A, B y C son acciones que hay que realizar y G es el objetivo deseado). No se han producido grandes logros en el camino hacia la inteligencia artificial general usando esos métodos, pero incluso los científicos de IA modernos como Stuart Russell continúan insistiendo en que la lógica simbólica será un componente importante de cualquier sistema de inteligencia artificial general en el futuro —ya que la inteligencia trata, entre otras cosas, el razonamiento y la planificación.

Así, Aristóteles dio el pistoletazo de salida a los estudios formales sobre la inferencia hace miles de años. Hace unas pocas décadas, también ayudó a dar el pistoletazo de salida al trabajo en la IA. El razonamiento simbólico que usa reglas deductivas enlaza la inteligencia específicamente con el conocimiento, un prerrequisito del sentido común, que es lo que sigue faltando de manera casi completa en los sistemas de IA. John McCarthy, pionero de los inicios de la IA como fundador de la disciplina durante la conferencia de Dartmouth de 1956, se dio cuenta muy pronto de ello, y dedicó un esfuerzo continuado al desarrollo de los sistemas de conocimiento —sistemas que dependen, para razonar y actuar, de proposiciones sobre el mundo que se pueden representar por ordenador—. Todos los sistemas de conocimiento antiguos se encontraron con problemas abrumadores, pero instructivos. Quizá pueda volverse sobre algunos de esos problemas con la esperanza de obtener algún avance. Otros problemas, no obstante, parecen fundamentales. En concreto, son limitaciones inherentes al mismo razonamiento basado en reglas. La lógica deductiva resulta precisa porque nos proporciona certezas. Tal y como cabría esperar, la certeza le pone el listón bastante alto al mundo real, donde los sistemas de inteligencia artificial general (y la gente) deben demostrar su inteligencia.

LA DEDUCCIÓN: CÓMO NO EQUIVOCARSE NUNCA

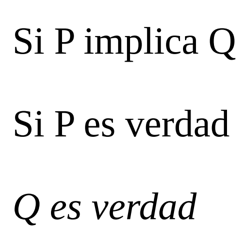
Los lógicos (y los científicos informáticos) analizan las inferencias deductivas con sistemas de proposiciones que pueden ser verdaderas o falsas. Dicta la convención que todas las proposiciones que preceden a la última dentro de una serie se llaman «premisas». La última proposición es una consecuencia de las premisas, y se conoce como «conclusión». Juntas, las premisas y la conclusión se denominan «argumento». Un buen argumento deductivo es una «apuesta segura», porque su conclusión es necesariamente verdadera. Aquí hay uno:

Cuando llueve, las calles se mojan.

De hecho, está lloviendo.

Por consiguiente, las calles están mojadas.

La conclusión es la inferencia que deberíamos extraer de las dos premisas. (En esencia, da respuesta a la pregunta: sin saber nada más, ¿qué se desprende de las premisas? La regla que se usa para inferir la conclusión es válida cuando la conclusión ha de ser verdadera siempre que las premisas lo sean también. La validez es un sello de «confianza» sobre esa regla, que preservará su veracidad siempre que nuestras premisas (o convicciones previas) sean verdaderas. Por consiguiente, el ejemplo de arriba es válido. Y utiliza una de las reglas deductivas más antiguas, que sigue recibiendo un nombre en latín: modus ponens. De forma casi simbólica:



Si P implica Q

Si P es verdad

Q es verdad

Y, en formato completamente analizable (computable), tenemos su forma lógica:

$P \rightarrow Q$

P

———

Q

Aquí, el conector «→» tiene un significado específico, o semántico, que determina el valor en cuanto verdad de P y de Q. En la lógica deductiva, la regla se conoce como «condicional material», y nos garantiza que Q se deriva del

carácter verdadero de P y de la regla $P \rightarrow Q$. (El rango de posibilidades verdaderas o falsas viene dado por una tabla de verdad, que mostraremos más tarde.)

Ahora pasa a considerar algunas modificaciones en nuestro argumento sobre la lluvia y las calles. En especial, ¿qué pasa si no llueve? En ese caso, la regla no «se dispara». No se deduce nada. Pero la forma del argumento sigue resultando válida. Sigue siendo verdad que, cuando llueve, las calles se mojan. Cuando de hecho llueve, el argumento se vuelve «sólido» (y no tan solo válido). La solidez es verdad; una verdad real, en oposición a la verdad condicional de la validez. La solidez nos indica que las premisas son realmente «verdaderas». La solidez garantiza que los agentes inteligentes que usen la inferencia deductiva inferirán verdades a partir de verdades anteriores. La validez, por otro lado, solo garantiza que, crea lo que crea el agente inteligente, sus inferencias serán correctas formalmente (incluso si razona acerca de mentiras o falsedades). De hecho, los argumentos deductivos que son válidos, pero no sólidos, pueden introducir todo tipo de ridiculeces en el razonamiento deductivo. Por ejemplo:

Cuando llueve, los cerdos echan a volar.

Está lloviendo.

Por consiguiente, los cerdos están volando.

Se trata de un argumento estúpido, pero perfectamente válido, porque de nuevo utiliza el modus ponens, el modo de razonar a partir de una proposición hipotética. Por supuesto, la primera premisa es falsa. La segunda premisa también podría ser falsa si no está lloviendo en realidad. No obstante, por mucho que llueva, no podemos fiarnos de la primera premisa, porque no existe ninguna conexión entre la lluvia y que los cerdos vuelen —y, de todos modos, los cerdos no vuelan, con independencia del tiempo que haga o de cualquier otra cosa—. El argumento es válido, pero no sólido... y es completamente inútil.

He aquí una deducción sólida:

Todos los hombres son mortales.

Sócrates es un hombre.

Por consiguiente, Sócrates es mortal.

¿Cómo podría estar equivocada? No puede. La conclusión siguiente siempre presenta un 100 % de certeza. La deducción proporciona un patrón para el pensamiento «perfecto» y preciso de los seres humanos y las máquinas, y es en buena medida por ese motivo por lo que ha sido investigada de manera extensiva en matemáticas y en las ciencias, y se ha utilizado con éxito en numerosas aplicaciones de importancia en el ámbito de la IA. En sus inicios, por ejemplo, los sistemas de IA basados en la deducción eran capaces de demostrar de manera automática teoremas reales (no «de juguete») en matemáticas. Un programa informático llamado Logic Theorist, creado por los pioneros de la IA Alan Newell, Herb Simon y Cliff Shaw, demostró teoremas lógicos de interés en una fecha tan temprana como 1956 sirviéndose de los Principia Mathematica de Bertrand Russell y Alfred North Whitehead, la obra fundacional de la lógica del siglo XX. Los sistemas de razonamiento automático que utilizan la deducción también se han aplicado al diseño de circuitos para placas base de ordenador, y a la tarea de verificar software y hardware, asegurándose de que el software no contiene errores ni contradicciones.[1] En tales casos, el enfoque deductivo resulta más sencillo y efectivo que los métodos modernos de la IA, que usan estadísticas y aprendizaje. Los primeros investigadores en IA sabían también que nuestro conocimiento se expresa a menudo de manera simbólica (como en el ejemplo de la lluvia), así que la deducción tiene sentido; es una elección obvia. Por desgracia, existen problemas bien conocidos a la hora de extender la inferencia deductiva a la inteligencia general.

PROBLEMAS DE CONOCIMIENTO

Con el paso de los años se han ido descubriendo numerosos problemas relacionados con la deducción. Quizá el más dañino haya sido que la deducción

nunca añade conocimiento. Si sé que la gente es mortal (se muere) y que tal y Pascual es una persona, ya sé que tal y Pascual se va a morir. La deducción tan solo confirma la conclusión a la que una persona racional debería haber llegado a partir de las premisas proporcionadas, cosa que es fácil de ver en un silogismo simple porque el «conocimiento» se hallaba ya contenido en las proposiciones. La conclusión se limita a hacerlo explícito.

La deducción resulta extraordinariamente útil como defensa contra la posibilidad de que alguien infiera conclusiones delirantes o incorrectas a partir de un conjunto de proposiciones —por ejemplo, que insistan en que, según las premisas de la mortalidad humana y el hecho de que Sócrates fue un ser humano, deberíamos concluir que Alfa Centauri está hecha de queso. La deducción otorga a los agentes racionales una plantilla para que no «se salgan de la senda», lo cual representa de manera evidente un buen primer paso para cualquier sistema de IA del que esperamos que acabe realizando inferencias inteligentes. Pero, usando solo la deducción, no llegaremos demasiado lejos. Por ejemplo, como respuesta a la teoría copernicana de que la Tierra gira alrededor del Sol y no al revés, los astrónomos de la vieja escuela ptolemaica podrían haber empleado un contraataque deductivo:

Si Dios creó el cielo, la Tierra se encontraría en el centro del cielo.

El cielo fue creado por Dios.

Por consiguiente, la Tierra se encuentra en el centro del cielo.

El argumento es válido pero, una vez más, esto solo nos indica que, si las premisas son de hecho verdaderas, la conclusión la sigue de manera necesaria. Todo el trabajo pesado recae en las preguntas empíricas acerca de la veracidad de las premisas. En la indagación sobre la mortalidad de Sócrates eso nos sale «gratis», por así decirlo, ya que en general todos coincidimos en que la gente se muere (por mucho que después vayan al cielo). Pero la generalización según la cual todo cielo creado por una divinidad tendría nuestro planeta en su centro parece tan debatible como cualquier otra afirmación bíblica o estética. Podríamos insistir en una interpretación alternativa de las Escrituras (es famosa la afirmación de Galileo de que Dios nos dice cómo funciona lo de ir al cielo, no

cómo funciona el cielo). O podríamos, sobre todo si somos ateos o materialistas científicos, rechazar la veracidad de la segunda premisa sin pensárnoslo demasiado.

Por tanto, la deducción se vuelve inútil para la búsqueda de nuevos conocimientos; solo sirve para esclarecer convicciones enfrentadas cuando se han cometido auténticos errores en el razonamiento. Como todo el mundo sabe, es posible que los teóricos de la conspiración nunca cometan errores de razonamiento deductivo; es solo que adoptan como verdaderas unas premisas que para otras personas son dudosas o simplemente delirantes.

En otras palabras, todo sistema inteligente requerirá otros tipos de inferencia para centrarse en unas convicciones verdaderas (y útiles). No basta con la certeza deductiva de una conclusión inferida.

PROBLEMAS DE RELEVANCIA

La deducción presenta otras limitaciones que la vuelven inadecuada como estrategia para diseñar la inteligencia general. Una especialmente dañina involucra los factores de relevancia. La premisa «Si está lloviendo, los cerdos volarán» es falsa, porque los cerdos no vuelan, pero también es un ejemplo excepcionalmente malo de cómo decir algo relevante. La lluvia no tiene nada que ver con la cuestión de que los cerdos vuelen. Por otro lado, los aviones sí que vuelan, pero la premisa «Si está lloviendo, los aviones volarán» también resulta irrelevante. Podría ser verdadera (al menos en algunos casos), pero el hecho de que llueva no debería conducir a que tengamos alguna convicción acerca de que los aviones se eleven por los aires. Una vez más, la proposición ignora los factores de relevancia.

Parte del problema tiene que ver con la causalidad: la lluvia no hace que los aviones vuelen (aunque, en algunas circunstancias, puede hacer que se queden en tierra). Aquí depende de la manera en que queramos usar el conocimiento. La proposición «Si el termómetro llega al color rojo, ahí fuera hace calor» es verdadera. Pero, si queremos inferir una explicación posible para una ola de calor, el termómetro no nos ayuda en nada. La proposición es verdadera pero

irrelevante. «Si el gallo canta es que está saliendo el sol» también es verdadera, pero si tuviéramos que preguntarle a un sistema de inteligencia artificial general por qué ha salido el sol y este nos saliera con el gallo, nos costaría atribuirle una gran inteligencia.

Piensa en este ejemplo, tomado del filósofo de la ciencia Wesley Salmon:

Los hombres que toman pastillas anticonceptivas con regularidad no se quedan embarazados.

Un hombre toma con regularidad las pastillas anticonceptivas de su esposa.

Por consiguiente, el hombre no se queda embarazado.[2]

De hecho, se trata de un argumento deductivo perfectamente sólido, que sigue el modus ponens con premisas verdaderas. Pero que el hombre no se quede embarazado no tiene nada que ver con las razones que se dan. Estas son irrelevantes, porque los hombres no se quedan embarazados de todos modos. El argumento no explica nada. Podemos imaginarnos un robot armado con una vasta base de datos y reglas razonando de esa manera, utilizando la deducción. En realidad, no hay nada per se erróneo, pero el robot no ha comprendido nada... No sabe lo que es relevante y lo que es una estupidez.

Piensa en este ejemplo, más sutil:

Todo aquel que ingiera treinta gramos de arsénico morirá antes de veinticuatro horas.

Jonas ingirió treinta gramos de arsénico a la hora t.

Jonas murió antes de que pasaran veinticuatro horas desde la hora t.

Se trata de un argumento deductivo perfectamente correcto, pero no serviría para explicar la muerte de Jones si, por ejemplo, este hubiera ingerido el arsénico a la hora t y hubiera fallecido en un accidente de tráfico (quizá cuando acudía veloz al hospital) antes de expirar por envenenamiento. Aquí, de nuevo, el argumento se basa en una deducción correcta, pero resulta irrelevante. No nos cuenta nada. Es hasta engañoso. La relevancia, en otras palabras, a menudo presupone el conocimiento de la causalidad, donde un hecho produce en realidad un resultado, o hace que suceda algo.

Otro motivo por el que la deducción cae constantemente víctima de los problemas de relevancia es que, de manera invariable, hay muchas causas posibles para que ocurra algo en nuestra experiencia cotidiana (y en la ciencia). Accidentes como que se estrelle un avión, por ejemplo, se pueden analizar en general señalando una causa próxima (cercana) y una causa distante (alejada) que expliquen el desastre en conjunto. Por ejemplo, las recientes tragedias protagonizadas por la compañía Boeing. Después de que dos aviones Boeing 737 Max se estrellaran en un plazo de seis meses en 2018, los investigadores descubrieron un fallo en el software del sistema de estabilización, el Sistema de Aumento de Características de Maniobra (MCAS en sus siglas inglesas). Al rediseñarse el antiguo Boeing 737-800 se generó espacio para unos motores de mayor tamaño, pero a cambio de colocarlos algo más adelantados y ligeramente por encima de las alas. Eso condujo a que la velocidad vertical de ascenso durante el despegue fuera más elevada, lo cual podía provocar, bajo ciertas condiciones, problemas de sustentación. Los fallos de sustentación son malos —en potencia, catastróficos—, así que se dotó a los nuevos Max de un MCAS para que, cuando fuera necesario, este dirigiera el morro del avión hacia abajo y evitara el problema. Por desgracia, esa corrección del morro podía hacer que el Max se precipitara contra el suelo. Y eso hizo el MCAS: arrebató el control del aparato a los pilotos en sendas tragedias que provocaron la muerte de 157 personas en Indonesia y de otras 189 en Etiopía.

La investigación consiguiente reveló fallas en el software que controlaba el MCAS, así que se identificó una causa próxima. Pero la investigación resultante de la anterior también recalcó el empeño con el que Boeing había puesto en servicio los Max a fin de competir en el ahorro de combustible que proporcionaban las aeronaves de Airbus, su gran rival —lo cual señaló una causa de fondo o distante—. También se descubrió que los pilotos del nuevo Max no habían recibido la formación adecuada. A ello contribuyó sin duda que, en el lanzamiento de mercadotecnia de la aeronave rediseñada, Boeing asegurara que

el Max no requería que los pilotos ya familiarizados con el 737-800 tuvieran que someterse a un costoso proceso de aprendizaje. Así, aquellos trágicos accidentes pueden atribuirse a una multiplicidad de causas. Inferir el motivo por el que los Boeing 737 Max se estrellaron implica considerar diversas causas posibles, y quizá no haya una que por sí sola explique las catástrofes.

La deducción no puede dialogar con estos escenarios propios del mundo real. Al obligar a que las inferencias sean verdaderas más allá de cualquier duda, la deducción invariablemente pasa por alto aquello que puede ser verdad en contextos donde la relevancia viene determinada por una mezcla de factores que no son necesarios, pero que sí se mantienen operativos en ciertas situaciones. En el universo platónico de formas inmutables, los triángulos han de tener tres lados y algunas cosas son Verdad, con uve mayúscula. En la experiencia desordenada del mundo, observamos o analizamos pocas cosas parecidas al triángulo. Son más bien como el Boeing 737 Max —o igual que una conversación común y corriente (como ya veremos)—. La inteligencia —sea lo que sea— va más allá de las deducciones. Nosotros mismos somos sistemas cognitivos, y es evidente que no somos solo sistemas deductivos. Eso sugiere que, para tener éxito, la IA de nivel humano tampoco puede ser completamente deductiva.

Tras el fracaso de lo que los críticos catalogaron como « inteligencia artificial de la vieja escuela», que dominó la IA antes de la era contemporánea (hasta bien entrada la década de 1990), los científicos de IA abandonaron los enfoques deductivos para ponerse a inferir en masa. En efecto, a numerosos lectores jóvenes les parecerá extraño que los practicantes de la disciplina se tomaran alguna vez en serio cosas como las «reglas» y los enfoques deductivos. Así fue. Pero las limitaciones devastadoras de la inferencia deductiva acabaron por condenar ese enfoque. Y, con la explosión de la red, las cantidades de datos disponibles para los llamados métodos superficiales o estadísticos hicieron que los sistemas deductivos o reglados parecieran menos útiles y más torpes. Un nuevo paradigma —un tipo de inferencia diferente— cobró prominencia en el trabajo serio sobre la IA. Se llama «inducción», y hablaremos de él a continuación.

EL PODER Y LOS LÍMITES DE LA INDUCCIÓN

La inducción implica la adquisición de conocimientos a partir de la experiencia. Por lo general, la experiencia es una interpretación en forma de observaciones —ver cosas—, aunque también puede provenir de cualquiera de nuestros cinco sentidos. (Tocar un fogón caliente es un ejemplo de inducción táctil.) A diferencia de la deducción, la forma general de la inducción pasa de las observaciones particulares a las hipótesis generales. La hipótesis inductiva cubre —esto es, explica— una observación. El mecanismo primario de la inducción es la enumeración: resulta difícil inducir los rasgos de una población de, pongamos, aves (por utilizar un ejemplo famoso) sin haber observado antes muchos ejemplos de aves. El carácter central de la enumeración desempeña un papel central en todas las versiones de la inducción, y será importante a la hora de comprender su naturaleza y sus limitaciones.

El poder de la inducción se debe no solo a que ayuda a organizar el mundo de las cosas en categorías a través de hipótesis (todos los objetos de X presentan la propiedad Y), sino que, además, otorga capacidades predictivas a los agentes que la utilizan.[3] Si cada vez que se acaba un partido de la liga de béisbol las calles del centro se llenan a reventar de gente, podría inferir que eso volverá a pasar la próxima vez que se acabe un partido —lo cual es una predicción—. La inducción atrapa la idea cotidiana de que, al observar lo que acontece en el mundo, obtenemos la capacidad para explicarlo y predecirlo. Muchas de nuestras expectativas se basan en la inducción. Si alguien desplazara el picaporte de la puerta de entrada de tu casa diez centímetros a la izquierda, lo más probable es que fallaras al ir a cogerlo. Tienes una teoría implícita —esto es, una hipótesis—, basada en los numerosos ejemplos previos en que lo has visto y cogido, y del lugar en el que se encuentra el picaporte.

La inducción tiene otras virtudes. Para comenzar es sintética, por tomar prestado el concepto de Kant; añade conocimiento. Puedo buscar en la red cuál es la hora punta del tráfico en la esquina de las calles Tercera y Mayor, pero si trabajo en el cruce de las calles Tercera y Mayor puedo mirar por la ventana. Esta última opción es una observación de primera mano que facilita mis inferencias inductivas y conforma expectativas y planes para cuando me vaya a casa. Por desgracia, la potente flexibilidad de la inducción (ligada a nuestros sentidos) también implica que no se trata de algo demostrable, ni de una verdad garantizada, como la deducción. El conocimiento cosechado por las observaciones es siempre provisional. ¿Por qué? Porque el mundo cambia. El futuro podría falsificar mis hipótesis inductivas. Es posible que mi coche se haya puesto en marcha mil veces sin problemas. Mañana por la mañana (cuando esté

llegando tarde a una reunión: la ley de Murphy) quizá no sea así. Eso es la inducción. El cambio llega (o no, por desgracia), y la observación previa por sí sola no puede indicarnos cómo ni cuándo lo hará.

La fortaleza de la inducción, no obstante, radica en el hecho de que la inteligencia está ligada de manera importante al mundo que nos rodea. La ciencia moderna sería imposible sin su lealtad a la inducción como método de conocimiento a través de la experiencia.

Piensa de nuevo en las enumeraciones. En su forma más sencilla, la inducción requiere tan solo de la enumeración de las observaciones previas para llegar a la conclusión general o norma (ley). He aquí un argumento:

N cisnes observados han sido de color blanco [donde N es un número alto].

Por consiguiente, todos los cisnes son blancos.

O:

Toda la vida que hemos visto está basada en el carbono.

Por consiguiente, toda la vida está basada en el carbono.

Tal y como sugieren estos ejemplos, la enumeración simple (es decir, el conteo) de los rasgos o propiedades de algo a menudo conforma la base de nuestras pretensiones de conocimiento sobre el objeto como clase. Así, los cisnes son simplemente esas aves de color blanco; la vida no es más que el fenómeno que se deriva del carbono. En la ciencia (y en la vida) también nos resulta útil contar una historia sobre el motivo por el que los cisnes podrían ser blancos, o por el que la vida podría tener una base de carbono, pero, en sentido estricto, las explicaciones con que se responde a los por qué quedan fuera del alcance de la inducción, la que enumera o la otra.

Sin embargo, es la simplicidad de la inducción lo que la vuelve tan útil como forma de inferencia. Cuanto más observo alguna propiedad en un objeto, más confianza tengo en que esa propiedad sea parte integral de ese objeto. Si no dejo de comprobar las pelotas de una bolsa y estas son siempre blancas, en algún momento comenzaré a confiar en una generalización del tipo «Todas las pelotas de esta bolsa son blancas». Pero, de nuevo, si no he comprobado hasta la última pelota, siempre existe la posibilidad de que mi inferencia inductiva sea errónea. La inducción resulta útil, pero no ofrece un conocimiento cierto.

He aquí otro tipo de generalización inductiva:

La proporción Q de una muestra de la población presenta la propiedad P.

Por consiguiente, la proporción Q de la población presenta la propiedad P.

La inducción entre una muestra y una población es bastante común en las investigaciones científicas, y a lo largo de los años se han desarrollado sofisticadas técnicas estadísticas para ayudar a que estas generalizaciones sean lo más sólidas y estén lo más libres de errores posible, dada la evidencia observable de la que se pueda disponer. En términos intuitivos, las generalizaciones inductivas también tienen sentido: si observo 75 bolas blancas y 25 bolas negras en una muestra, a falta de otra evidencia, debo esperar que haya 750 bolas blancas en una población de 1.000. La inferencia parece correcta, lo que pasa es que no existe una certeza sobre ella.

El muestreo aleatorio también se basa en una generalización debida a las observaciones. Pruébalo por ti mismo: lanza una moneda al aire varias veces y cuenta el número de caras y de cruces que te salen. Eso es un muestreo aleatorio (puesto que no puedes darle ningún sesgo al lanzamiento, la moneda es justa). Es posible que te salgan dos o tres caras seguidas. Aunque resulte muy improbable, puede ser incluso que te salgan cinco caras o cinco cruces seguidas. Pero, dado un muestreo lo bastante amplio, podrás generalizar diciendo que las posibilidades de que salga cara o cruz son mitad y mitad. De ahí que la generalización inductiva sea «La moneda lanzada al aire saldrá cara quinientas veces de cada mil», lo cual se acerca bastante a la verdad. (La ley de los grandes números nos dice que, dado un muestreo lo bastante amplio, la probabilidad se

aproximará a la probabilidad real: al cabo de un millón de lanzamientos, estaremos bastante cerca de ese 50-50 %.) He aquí otro ejemplo popular de generalización estadística a través de la inducción:

El 73 % de un muestreo aleatorio de votantes apoya al candidato X.

Por consiguiente, el candidato X obtendrá cerca del 73 % de los votos.

Existe la posibilidad de que el candidato X se vea envuelto en un escándalo antes de las elecciones, lo cual invalidaría esa inferencia inductiva. Pero, una vez más, en ausencia de un conocimiento mayor, podemos razonar de esta manera y extraer conclusiones sobre lo que esperamos que suceda.

La IA moderna se basa en análisis estadísticos, y por tanto depende de un marco inductivo, lo cual resulta de utilidad para numerosas aplicaciones comerciales. Por ejemplo, la IA puede ofrecer recomendaciones: un tipo de predicción basado en observaciones pasadas. He aquí otro ejemplo, que resultará familiar a cualquiera que tenga un buscador de contenido.

El 75 % de las noticias que lee el usuario X son crónicas políticas de carácter conservador en el sitio web C.

Por consiguiente, el usuario X querrá leer esta noticia en C.

Es posible que al usuario X también le guste de vez en cuando leer algún artículo del New Republic. Por desgracia, lo más probable es que el sistema que infiere las preferencias de X sirviéndose de la inducción lo ignore. Se trata de un inconveniente evidente a la hora de fiarse de las generalizaciones inductivas derivadas de la observación —son sustitutivos de un conocimiento más profundo (lo que es peor, tienden a esperar que el futuro tenga el mismo aspecto que el pasado).

David Hume, filósofo del siglo XVIII y el primer pensador que señaló los límites de la inducción, aportó a filósofos y científicos lo que ahora se conoce como el problema de la inducción. Tal y como él mismo lo expresó, confiar en la inducción requiere de nosotros la convicción de que «los casos de los que no tenemos experiencia se parezcan a aquellos de los que sí la tenemos». En otras palabras, la regla general inductiva que aplicamos tiene que ser ampliable a ejemplos que no hemos visto, y no existe ninguna garantía de que vaya a mantenerse. A diferencia de la deducción, en la estructura de la inducción no hay nada que nos proporcione una certeza lógica. La inducción se limita a calcular que el mundo cuenta con ciertas características, y nosotros podemos examinarlo y sonsacarle el conocimiento (que creemos tener) sobre él.[4]

El problema de la inducción puede parecer una de esas preocupaciones menores a las que los filósofos se entregan por placer, pero en realidad los límites de la inferencia inductiva generan problemas constantes a los científicos en su búsqueda de teorías verdaderas. Hay ejemplos por doquier. Solíamos comernos solo la clara de los huevos porque la ciencia nutricionista nos había advertido contra los peligros de las grasas saturadas que se encontraban en la yema. Si pasamos algunas décadas a cámara rápida descubrirás que los científicos nutricionistas ahora nos animan a comer huevos con yema y todo, ya que ayudan a quemar grasas y levantan el ánimo; incluso protegen contra los problemas de corazón (el mismo motivo por el que nos provocaban ansiedad hace unas décadas). En un sentido muy real, podemos culpar a la inducción por esos cambios radicales tan vergonzosos. Estos suceden porque nuestras observaciones y comprobaciones nunca son completas. Las correlaciones pueden sugerir una causa subyacente en la que podamos confiar (un fragmento de conocimiento real), pero existe la posibilidad de que se nos haya pasado algo por alto mientras comprobábamos y observábamos las cosas que las afectaban. La correlación puede ser falsa o accidental. Podríamos haber estado buscando algo equivocado. La muestra podría haber sido demasiado pequeña o no representativa por razones que solo se vuelven aparentes más tarde. Es un problema habitual, y en el fondo se debe al espectro de la inducción y sus límites —al final resultó que los filósofos no nos estaban haciendo perder el tiempo.

En el fondo, toda inducción se basa en una enumeración. Quizá suene sospechosamente simple (o debería ser así): ¿es posible que, a fin de obtener teorías sobre el mundo, tengamos que limitarnos a contar ejemplos? En un sentido importante, sí. Una sola experiencia no permite la licencia de realizar inferencias inductivas. Si veo un orangután y conozco su aspecto, puedo

clasificarlo. Pero, si aún no conozco qué animal es, tendré que observar muchos ejemplares antes de averiguar si el animal que he visto es un chimpancé peculiar o la cría de un Big Foot adulto. Tal y como dijo Hume, para inferir las causas necesitamos ver «correlaciones constantes», y para inferir categorías o tipos necesitamos contar con ejemplos enumerados. (Como veremos, así es exactamente como funciona el aprendizaje automático.)

Por supuesto, el razonamiento inductivo se vuelve más complicado: en terrenos como la economía o las ciencias sociales, las inferencias estadísticas también son inductivas, pero para comprenderlas hay que saber muchísimo sobre teoría de probabilidades (y sobre economía y ciencias sociales). Y, en las ciencias, las nuevas inferencias inductivas se elevan de manera inevitable sobre otras más antiguas que los científicos han pasado a considerar sólidas y verdaderas. (Así que también tenemos que saberlo todo sobre esas teorías.) Pero, en el fondo, la inducción simplemente generaliza a partir de la observación de ejemplos. Cuando el origen de esas generalizaciones se puede explicar con una historia, con una causa o con un conjunto de causas, podemos confiar en que se haya adquirido un nuevo conocimiento..., aunque no sea cierto por necesidad, como con la deducción. Este cuenta con el apoyo de la observación y la comprobación.

La crítica que Hume realizó a la inducción fue sobre todo una crítica a la causalidad. La inducción no requiere que se conozcan las causas (de otro modo, no sería enumerativa). Si sabemos, por ejemplo, que el color del plumaje de las aves viene determinado en parte por las características de su hábitat, aunque todos los cisnes de Inglaterra sean blancos podemos esperar que se encuentren plumajes negros en cisnes de hábitats diferentes. Pero, a falta de una teoría, la inducción solo podrá indicarnos eso si nos ponemos a volar por todo el mundo y vamos observando a los cisnes en los diferentes lugares en los que viven. Las hipótesis que citan causas específicas son el objetivo de la observación, pero por desgracia los medios lógicos de la inducción no son adecuados para proporcionarlas. Hacen falta inferencias adicionales (y ahí la deducción puede ayudar, pero solo en parte).

La cuestión es la siguiente: la inducción entendida debidamente dentro del marco lógico de la inferencia es, aunque necesaria y habitual, bastante limitada. Y también suele ser malinterpretada, lo cual contribuye a un exceso de confianza general en el hecho de que la inducción garantice un conocimiento «científico» y de solidez empírica que nos libre de especulaciones rocambolescas. Sherlock Holmes, nuestro héroe detective, explica a veces su método como una serie de

inducciones meticulosas, observaciones simples y claras en las que no se entromete ninguna opinión, ni idea, ni convicción. Asegura, ante un perplejo y fascinado Watson, que él se limita a «observar las cosas con detenimiento». Holmes conoce el valor de la simple observación —cuanto más simple, mejor—, porque aquello que creemos saber puede impedir que veamos algo nuevo. Pero esa es solo una parte de la historia de la inteligencia. También tenemos que entender el sentido de lo que observamos. Holmes, igual que Dupin, soluciona crímenes recopilando observaciones de una manera novedosa. El diablo está en los detalles y los detalles se encuentran en la novedad, que no es ninguna inducción.

La inferencia inductiva nos sitúa inevitablemente delante de otro peligro, que el ojo crítico de Hume volvió memorable: los hechos recién descubiertos pueden sorprendernos. En entornos dinámicos como la vida cotidiana, la observación es abierta. Observaciones futuras pueden revelar lo que antes se encontraba oculto o nos resultaba desconocido: ¡sorpresa! Y nuestra propia confianza en la inferencia inductiva puede llevar a que nos cueste más evitar sus inevitables defectos y fracasos. Eso nos lleva a hablar de las fiestas navideñas o, al menos, del «pavo inductivo» de Bertrand Russell.

EL PAVO DE RUSSELL

Bertrand Russell fue uno de los filósofos e intelectuales públicos más famosos del siglo XX. Un lógico, matemático y activista social que en una ocasión pasó seis meses en la cárcel por protestar contra la entrada de Gran Bretaña en la primera guerra mundial. Más tarde, en los años cincuenta protestó contra la proliferación de armas nucleares. Sus intereses intelectuales también fueron una forma de protesta: le preocupaba que se pudiera usar el lenguaje para inventarse problemas y soluciones filosóficas, y pensó que el antídoto para esa filosofía ensoñada consistía en ligarla a los métodos de la ciencia.

Pero, tal y como el propio Russell señaló, la ciencia a menudo avanza sin unas reglas inferenciales claras. Para exponer la inferencia en la ciencia, por tanto, debemos revelar los errores de nuestro pensamiento acerca de la investigación científica y la búsqueda de la verdad en general. Ese fue el motivo por el que

decidió poner la lupa sobre el problema de la inducción, que definió como uno de los «problemas filosóficos» fundamentales en un libro titulado precisamente Los problemas de la filosofía, y argumentó, igual que sir Karl Popper, que la ciencia no acumula conocimiento coleccionando o enumerando hechos. En otras palabras, que no obtenemos conocimiento científico a través de la mera inducción. De hecho, la inducción en sí misma es irremediablemente defectuosa.

Russell ofreció un ejemplo evidente y accesible: que observemos el sol elevarse cada mañana no nos ofrece ninguna prueba de que vaya a hacerlo de nuevo. Nuestra confianza en que el sol vaya a salir mañana no es más que un «hábito de asociación», tal y como lo denominó Hume. No es solo que la inducción resulte incompleta, sino que con toda seguridad no puede confirmar teorías científicas ni convicciones desde la enumeración de observaciones. Nuestra creencia de que sí lo hace genera todo tipo de distorsiones. La «falacia del jugador», por ejemplo, es una convicción tozuda que aparece en los ludópatas, según la cual la frecuencia pasada de un resultado nos comunica algo verdadero acerca de resultados futuros. La falacia puede apoyar la expectativa de que siga habiendo más de lo mismo o lo opuesto: que ha llegado el momento de que pase algo nuevo. Las rachas, cuando se juega a los dados, generan la ilusión de que de algún modo habrá una influencia sobre la siguiente tirada: la racha de buena suerte se mantendrá así (estoy que me salgo) o la racha de mala suerte tendrá que acabar (ya me toca). Cualquiera de los dos escenarios puede producirse, por supuesto, pero la moraleja importante es que la siguiente tirada de los dados es independiente de todas las tiradas anteriores. La buena racha continúa si los dados caen aleatoriamente de una manera y se rompe si caen de otra. Este es un ejemplo de nuestra ansiedad por aplicar la falacia inductiva incluso a los hechos más aleatorios.

Sin embargo, la mayor parte del mundo real no es aleatoria, y eso hace que cueste aún más arrancar de raíz la tendencia a ver patrones inductivos incorrectos —la verdad es que los patrones están «ahí fuera», pero no siempre podemos averiguar cuáles son los verdaderos solo a través de la observación—. Vemos regularidades y patrones por todas partes. Más allá de la ludopatía, este peculiar giro mental ayuda a explicar nuestra inclinación a generalizar a partir de la observación. Los cisnes son blancos. El sol volverá a salir. El ascensor siempre me está esperando en la planta baja a las 3:30 de la mañana. Por supuesto que hay generalizaciones en las que se puede confiar —las vemos por todas partes, y no es ilusorio hacerlo—, pero el problema de la inducción, tal y como señaló Russell, es que no tenemos ninguna base para inferir conocimientos

basándonos solo en esas generalizaciones. La ciencia debe confiar en estrategias inferenciales más profundas y poderosas. La inducción misma es fina como el papel.

Russell nos ofrece, a modo de ejemplo de los límites de la inducción, un ave de corral bien alimentada que resulta ser una fabulosa pensadora inductiva. He aquí una versión de su triste historia:

El pavo descubrió que, aquella primera mañana en la granja de pavos, le dieron de comer a las 9. No obstante, puesto que era un buen inductivo, no sacó conclusiones precipitadas. Esperó hasta haber reunido un amplio número de observaciones del hecho de que le daban de comer a las 9 de la mañana, y realizó esas observaciones bajo una amplia variedad de circunstancias, los miércoles y los jueves, los días de calor y los días de frío, los días de lluvia y los días secos. A diario, añadía otra proposición observacional a su lista. Al fin, su conciencia inductiva se dio por satisfecha y elaboró una inferencia inductiva para concluir: «Siempre me dan de comer a las 9 de la mañana». Pero, ay, esa conclusión se reveló falsa con una certeza absoluta cuando, el 24 de diciembre, en vez de darle de comer le cortaron el cuello. Una inferencia inductiva con premisas verdaderas ha conducido a una conclusión falsa.[5]

El pavo de Russell revela el disparate de crear «hábitos de asociación» sin tener un conocimiento más profundo de las regularidades que observamos. Pero el conocimiento a menudo es una convicción disfrazada: aquello que creemos saber puede ser erróneo.

Un segundo problema igual de dañino de la dependencia en la inferencia deductiva tiene que ver con la ausencia de conocimiento. Buena parte del mundo se halla oculta de manera misteriosa; está enfangada en lo aleatorio y lo caótico, o simplemente es demasiado compleja para que confiemos solo en la inducción. Me vienen a la cabeza los mercados financieros. Podemos intentar predecir el rendimiento de un paquete de acciones con todo tipo de técnicas sofisticadas, pero como cualquier agente de bolsa sabe bien, los resultados del pasado no son indicativos de los resultados del futuro. Y, si hemos de ser sinceros, buena parte de nuestra experiencia del mundo presenta esa cualidad frustrante. Sabemos que

el ascensor se queda en la planta baja cuando no lo usa nadie y, por inducción, es posible que creamos que esté allí esperándonos si volvemos a casa pronto del trabajo, porque estaremos fuera del horario del resto de la gente. Pero alguien podría estar mudándose, o unos parientes de Minnesota han venido a visitar a fulano de tal, etc. Las reglas están hechas para romperse, y las expectativas, también.

Nuestras predicciones se ven constantemente frustradas porque el conocimiento que necesitamos para amplificar la inducción a menudo no existe o no se puede disponer de él. Puedo ver mil cisnes blancos en Inglaterra y llegar a la conclusión de que «Todos los cisnes son blancos». Ese mismo año, durante un viaje por Australia, veo un cisne negro... y la inducción se va al garete. Buena parte de lo que creemos saber en realidad es provisional, está pendiente de exámenes ulteriores, y es esa sobredependencia de la inducción la que lleva a que los cambios parezcan sorprendentes. En las ciudades grandes del oeste de Estados Unidos, como Seattle, los conductores suelen reducir la velocidad o detenerse ante un semáforo en amarillo en vez de pisar a fondo el acelerador para dejarlo atrás. También tratan con deferencia a los peatones, en vez de sortearlos. Eso podría hacer que me sorprendiera el comportamiento de los conductores de Nueva York o de Bombay. Incluso bajo una misma normativa, la conducta difiere. Si me fío de los datos e inducciones de mis experiencias pasadas, es posible que me choquen por detrás o que me peguen un bocinazo.

Así que ¿por qué debería confiar únicamente en hechos pasados, con toda esta nueva información? ¿Qué sería lo más razonable e inteligente por mi parte? ¿Cómo debería tratar la nueva información? ¿Y si veo una tira de clavos atravesando la carretera, cosa a la que nunca me había enfrentado, o una fila de patos cruzándola, o señales de tráfico con las que no estoy familiarizado? Por desgracia, la respuesta en estos casos no es una mayor inducción, sino menos.

LA INDUCCIÓN FUNCIONA EN LOS JUEGOS, NO EN LA VIDA

El mundo real es un entorno dinámico, lo que quiere decir que se encuentra en cambio constante de maneras tanto predecibles como impredecibles, y que no

podemos acotarlo con un sistema de reglas. Los juegos de mesa, no obstante, sí están acotados por sistemas de reglas, lo cual contribuye a explicar por qué los enfoques inductivos que aprenden a partir de la experiencia del juego funcionan tan bien. AlphaGo (o su sucesor, AlphaZero) utiliza un tipo de aprendizaje automático conocido como aprendizaje profundo para jugar al go, un juego de gran dificultad. Juega contra sí mismo, usando algo que se llama aprendizaje profundo por refuerzo, e induce hipótesis acerca de los mejores movimientos que se pueden realizar sobre el tablero dada su posición y la del rival. El enfoque ha tenido un éxito fabuloso en «juegos de reglas conocidas, con dos participantes, discretos y observables», tal y como señala el científico de IA Stuart Russell.[6] Es posible que Russell no haya estado pensando en el pavo del otro Russell, pero debería haberlo hecho: el problema real con los juegos que apuntalan la IA es que permiten la formación de hipótesis (generalizaciones nacidas de la experiencia) según unas reglas conocidas. Irónicamente, igual que la IA clásica antes, esas reglas no tienen aplicación en el mundo real, lo cual representa el quid de la búsqueda de una inteligencia general.

Los científicos informáticos que confían en los métodos inductivos suelen desdeñar y catalogar de irrelevantes los problemas de inducción señalados por Hume (o por Russell). Según la lógica, por supuesto que no existen garantías de corrección cuando se usa la inducción, pero sí que podemos «acercarnos bastante» a ella.

Esa respuesta yerra por completo el tiro. Un método catalogado como «probable, aproximadamente correcto» rige la formación de hipótesis para la IA estadística, como el aprendizaje automático, y se sabe que resulta efectivo para extirpar, con el tiempo, aquellas hipótesis que sean malas o falsas. Pero en realidad ese método es una extensión del argumento original de Hume, según el cual la inducción no ofrece ninguna garantía de corrección más que cuando se aplica a escenarios como el de los juegos, que disponen de reglas con las que encerrar las inferencias estadísticas. Una solución probable y aproximadamente correcta no altera el problema de la inducción en entornos dinámicos y ajenos al mundo de los juegos o a los laboratorios de investigación.

Los investigadores de IA son conscientes del problema de la inducción (explícita o implícitamente), pero este rara vez entra en las críticas al aprendizaje automático (o al aprendizaje profundo) porque, en esencia, estas se dedican a darle la vuelta al problema. Puesto que la inducción no funciona en entornos dinámicos, aceptan, la aplicamos en entornos controlables. Lo cual es como

buscar las llaves debajo de una farola porque allí hay más luz que donde se han caído. Es cierto que los seres humanos han «solucionado» el problema de la inducción lo bastante bien como para utilizar la experiencia de manera efectiva en el mundo real (¿y dónde, si no?). Pero hemos solucionado ese problema no con una forma más potente de inferencia inductiva, sino combinándola de algún modo con tipos más potentes de inferencia que contribuyan a la comprensión. El aprendizaje automático se da solo con inducción (tal y como comentaremos en el capítulo 11), así que los investigadores de la disciplina deberían mostrarse más escépticos de lo que son habitualmente sobre las perspectivas de la inteligencia artificial general.

REGULARIDAD Y FRAGILIDAD

La inducción permite a la inteligencia comportarse como un detector de regularidad. La IA estadística destaca a la hora de capturar regularidades a partir del análisis de datos, y ese es el motivo por el que las labores de reconocimiento visual de objetos, como la identificación de fotos de rostros humanos y mascotas, se cuentan entre sus éxitos. Los píxeles de un rostro se pueden distribuir y regular de manera que sean estudiados y clasificados. Sin embargo, puesto que esos sistemas aprenden a partir de las observaciones de unos patrones de impulso específicos, tienen problemas de fragilidad. Tal y como han señalado Gary Marcus, Ernest Davis y otros investigadores, incluso cambios en apariencia benignos como pasar el color de fondo del blanco al azul en las tareas de detección de objetos pueden llevar a que su rendimiento se degrade. Atestar las fotos con otras imágenes también tiene como resultado una degradación severa.[7] La persona ignorará sin problemas que se añadan unas letras sin importancia en la zona roja de una señal de stop, pero cuando le presenten esta imagen alterada a un sistema de aprendizaje profundo, este la catalogará como una señal de límite de velocidad. Y hay ejemplos parecidos en el mundo real, incluyendo sistemas autónomos de navegación en coches sin conductor que han catalogado de manera errónea un autobús escolar como una máquina quitanieves, y un camión que giraba como un paso elevado.

El aprendizaje automático es inductivo porque adquiere conocimientos a partir de la observación de los datos. La técnica conocida como aprendizaje profundo

es un tipo de aprendizaje automático —una red neuronal— que se ha revelado muy prometedor a la hora de reconocer objetos en fotos, de estimular el rendimiento de los vehículos autónomos y de participar en juegos en apariencia difíciles. Por ejemplo, el sistema DeepMind de Google aprendió a jugar a diversos videojuegos clásicos de Atari y se celebró a bombo y platillo. Se proclamó que la suya era una inteligencia general, porque el mismo sistema era capaz de dominar diferentes juegos usando el llamado enfoque de aprendizaje por refuerzo profundo que propulsaba AlphaGo y AlphaZero. Pero la empresa emergente de IA Vicarious, por su parte, no tardó en señalar que cualquier cambio en apariencia inocuo sobre los juegos degradaba de manera manifiesta la actuación fabulosa del sistema. En Breakout, por ejemplo, el jugador mueve una pala de lado a lado sobre una línea base y va golpeando una pelota para lanzarla hacia arriba y contra un muro de varias capas de ladrillo. Con cada golpe destruye un ladrillo (y se acerca un poco más al «Breakout», a la evasión), pero cuando la bola rebota el jugador debe preocuparse de que no se le escape. Acercar la pala unos pocos píxeles a los ladrillos tuvo como resultado una degradación severa de su desempeño. «El sistema entero de DeepMind se derrumba», observaron Marcus y Davis en su crítica a la IA moderna. Y citaron una observación del pionero de la IA Yoshua Bengio según la cual las redes neuronales profundas «tienden a aprender regularidades estadísticas en el conjunto de datos en vez de conceptos abstractos de nivel superior».[8]

Algo que se suele ignorar o malinterpretar es que esos fracasos son fundamentales y no se pueden remendar con enfoques de aprendizaje más potentes y dependientes de la inferencia inductiva (basada en datos u observaciones). El problema aquí es el tipo de inferencia, no las especificidades de un algoritmo. Puesto que se necesitan muchos ejemplos para estimular el aprendizaje (en el caso del go, los juegos de ejemplo se elevan a varios millones), los sistemas son motores de inducción enumerativa glorificados, guiados por la formación de hipótesis dentro de los límites de las características del juego y de sus reglas. Esos mundos están cercados por reglas y son regulares —se trata de una especie de mundo en forma de curva de campana donde los mejores movimientos son los que con mayor frecuencia conducen a la victoria—. No se trata de ese mundo real que la inteligencia artificial general debe dominar, y que descansa más allá de los juegos de diseño humano y los centros de investigación. Esa diferencia lo es todo.

El pensamiento del mundo real depende de la detección sensitiva de la anormalidad o de las excepciones. Una calle urbana concurrida, por ejemplo,

está llena de excepciones. Es uno de los motivos por los que no hay robots paseándose por Manhattan (ni, por otro motivo relacionado con las excepciones, conversando con los seres humanos). Un robot manhattiano no tardaría en tropezarse, provocaría un atasco de tráfico al aventurarse de manera desaconsejable a cruzar la calle, chocaría contra la gente o algo peor. Manhattan no es Atari ni el Go, y tampoco es una versión a escala de ellos—. Un «cerebro» de aprendizaje profundo sería (y es) un grave lastre en el mundo real, igual que cualquier sistema inductivo que pretenda reemplazar la inteligencia genuina. Si pudiéramos enseñar al pavo de Russell a que «jugara» a no convertirse en la cena, quizá aprendería a esfumarse el día antes de Navidad. Pero en ese caso no sería un buen pavo inductivista; dispondría de un conocimiento previo, suministrado por los seres humanos.

La IA estadística acaba así con un «problema de larga cola», donde los patrones habituales (en la cola gruesa de la curva de distribución) resultan sencillos, pero los raros (su cola larga) son difíciles. Por desgracia, algunas de las inferencias que realicen los sistemas de IA de inteligencia humana se encontrarán en la cola larga, no en el dulce punto de inducción de las regularidades localizables en los sistemas de mundo cerrado. De hecho, al centrarse en los éxitos «fáciles» que explotan las regularidades, la investigación en IA corre el peligro de alejarse de manera colectiva del avance hacia una inteligencia general. Ni siquiera estamos realizando un progreso gradual, porque, en la práctica, trabajar los problemas sencillos implica descuidar los reales (las llaves no están cerca de las farolas). Por sí mismas, las estrategias inductivas generan una esperanza falsa.

Que una foto quede mal catalogada en Facebook o que Netflix nos recomiende una película aburrida quizá no represente un problema demasiado grande con la inducción dirigida por datos, pero los coches sin conductor y otras tecnologías de carácter más crítico sin duda pueden meternos en un lío. Oren Etzioni, responsable del Instituto Allen de Inteligencia Artificial, afirma que el aprendizaje automático y los macrodatos son «modelos estadísticos de alta capacidad».[9] Eso habla de unas ciencias informáticas impresionantes, pero no de una inteligencia general. Las mentes inteligentes suman la comprensión a los datos, y pueden encontrar el sentido que lleve a apreciar los puntos de falla y las anormalidades. Los datos y el análisis de datos no son suficientes.

EL PROBLEMA DE LA INFERENCIA EN TÉRMINOS DE

CONFIANZA

En su esclarecedora crítica a la inducción que se utiliza en las predicciones financieras, el ex agente de bolsa Nassim Nicholas Taleb divide los problemas de predicción estadística en cuatro cuadrantes, siendo sus variantes: primero, si la decisión que se ha de tomar es simple (binaria) o compleja, y, segundo, si la aleatoriedad del asunto es «mediocre» o extrema. Los problemas del primer cuadrante reclaman decisiones simples en relación con una distribución de probabilidad de cola fina. Los resultados son relativamente fáciles de predecir en términos estadísticos, y los sucesos anómalos, cuando se dan, tienen un impacto pequeño. Los problemas del segundo cuadrante son fáciles de predecir, pero, cuando sucede algo inesperado, las consecuencias son importantes. Los problemas del tercer cuadrante implican decisiones complejas, pero consecuencias manejables. Y luego están los problemas «pavo», en el cuarto cuadrante, que implican decisiones complejas emparejadas con distribuciones de probabilidad de cola gruesa y consecuencias de alto impacto. Piensa en los cracs del mercado de valores. Taleb apunta al exceso de confianza en la inducción como uno de los factores clave que han exacerbado el impacto de esos sucesos. No se trata solo de que nuestros métodos inductivos no funcionen, sino de que, al fiarnos de ellos, no logramos utilizar enfoques mejores, y eso tiene consecuencias potencialmente catastróficas. En efecto, nos obsesionamos con el pensamiento maquinal, cuando el análisis del pasado no nos ayuda en nada. Este es uno de los motivos por los que la superinteligencia inductiva llevará a resultados estúpidos. Tal y como dice Taleb en broma, es importante saber cómo «no convertirse en un pavo».[10]

Por supuesto, la predicción cuenta con otros límites que no se pueden resumir con pulcritud desnudando los puntos ciegos de la inducción. Los cisnes negros, al fin y al cabo, son raros, igual que los cracs del mercado de valores y las guerras (y las innovaciones) de importancia. Se nos puede perdonar que usemos la inducción para ayudar a iluminar unas posibilidades que de todos modos son opacas y muy impredecibles, pero no que intentemos reemplazar nuestro entendimiento exclusivamente con datos y estadísticas. En algunos casos, como en los sistemas naturales caóticos (pongamos, los sistemas con turbulencia), ahora sabemos que habrá limitaciones inherentes a la predictibilidad cuando se use cualquier tipo conocido de método inferencial. La inducción puede sugerir que el pasado será parecido al futuro, pero la teoría del caos nos dice que no será

así —o, al menos, que no podemos determinar en qué sentido se parecerán—. En algunos casos, aunque incompleto, el análisis estadístico es todo lo que tenemos.

CAUSA PROBABLE

El ganador del premio Turing Judea Pearl, un célebre científico informático que ha dedicado su vida al desarrollo de métodos computacionales que se muestren efectivos con el razonamiento causal, sostiene en su obra de 2018 El libro del porqué que el aprendizaje automático nunca podrá ofrecernos una comprensión real porque el análisis de los datos no abarca el conocimiento de la estructura causal del mundo real, esencial para la inteligencia. La «escalera de causalidad», tal y como la llama él, parte de la asociación de puntos de datos (ver y observar), y continúa con una intervención en el mundo (hacer), lo cual requiere el conocimiento de las causas. A continuación, sigue avanzando hacia el pensamiento contrafactual, como la imaginación, el entendimiento y las preguntas de tipo: ¿y si hubiera hecho algo de manera diferente?

Los sistemas de IA que usan métodos de aprendizaje automático —y muchos animales— se encuentran en el peldaño más bajo, el de la asociación. En ese primer nivel, buscamos regularidades en nuestras observaciones. Es lo que hace la lechuza cuando observa los movimientos de la rata y averigua el lugar donde es probable que esté el roedor un momento después, y es lo que hace un programa de ordenador para jugar al go cuando estudia una base de datos de millones de partidas para averiguar qué movimientos están asociados a un porcentaje mayor de victorias.[11]

Aquí, Pearl nos hace el favor de relacionar observaciones y datos.[12] También señala que el ascenso por esa escalera implica diferentes tipos de pensamiento (o, de manera más específica, de inferencia). La asociación no guarda proporción con el pensamiento causal, ni con las imaginaciones. Podemos reestructurar el problema del salto entre la inteligencia artificial y la inteligencia artificial general como, precisamente, el problema de descubrir nuevas teorías que permitan subir por la escalera de Pearl (o, en el marco contemporáneo, el problema de pasar de la inducción a otros tipos de inferencia más potentes).[13]

UN MANUAL BÁSICO DE SENTIDO COMÚN

Es posible que tus padres, o tu pareja o un amigo, te hayan acusado alguna vez de no tener sentido común, pero anímate: tienes mucho más que cualquier sistema de IA, pero de lejos. Como bien sabía Turing, el sentido común es lo que permite que dos personas mantengan una conversación cotidiana. El problema del sentido común y de la comprensión de un lenguaje particular, que lo requiere, ha sido un motivo de inquietud notable entre los investigadores de IA desde la aparición de la disciplina. Y al final se está poniendo de manifiesto que el bombo y platillo que han acompañado el aprendizaje automático no nos están acercando nada a él. Los investigadores están reconociendo cada vez más este hecho, y no podría haber llegado en un mejor momento. Marcus y Davis se preguntan, si los ordenadores son tan listos, ¿por qué no pueden leer ni conducirnos por «El sentido común y el camino hacia el entendimiento profundo»? (uno de los capítulos de su libro).[14] Stuart Russell encabeza su listado de «Avances conceptuales que están por llegar» con los aún misteriosos «lenguaje y sentido común».[15] Pearl también entiende que la comprensión del lenguaje sigue sin resolverse (y ofrece su propio «minitest de Turing», que requiere el entendimiento de la causalidad).[16]

Así que, para progresar en la IA, tenemos que ir más allá de la inducción. (Si en una escalera metafórica te encuentras en el peldaño de la asociación, mira hacia arriba.) Hagámoslo a continuación —o al menos comencemos a hacerlo—. En el camino hacia la necesidad de una inferencia abductiva, primero deberíamos ir a lo específico; en particular, al aprendizaje automático y su fuente de entrada, el big data.

Capítulo 11

El aprendizaje automático y el big data

Aprender consiste en «mejorar unas prestaciones basándose en la experiencia».[1] El aprendizaje automático consiste en lograr que los ordenadores mejoren sus prestaciones basándose en la experiencia.

Esta definición del subcampo de la IA conocido como aprendizaje automático goza de una amplia aceptación y no resulta especialmente controvertida. Se ha mantenido esencialmente inalterada desde los trabajos tempranos sobre algoritmos de aprendizaje en los albores de la disciplina. Tom Mitchell, científico informático de Carnegie Mellon e investigador sobre el aprendizaje automático desde hace mucho, ofreció una definición ligeramente más detallada en su obra Machine Learning [«Aprendizaje automático»], de 1997: «Se dice que un programa informático ha aprendido de la experiencia E con respecto a algún tipo de tareas T y una medición del rendimiento P cuando su rendimiento con las tareas de T, medido por P, mejora con la experiencia E».[2] En otras palabras, el aprendizaje automático es el tratamiento informático de la inducción: la adquisición de conocimiento a partir de la experiencia. El aprendizaje automático no es más que una inducción automatizada, así que no debería sorprendernos que los problemas con la inferencia inductiva impliquen otros problemas para el aprendizaje automático. Desarrollar esos problemas inevitables es el objetivo de este capítulo.

Hay dos tipos principales de aprendizaje. Cuando los seres humanos etiquetan el dato de entrada para señalar el resultado deseado, se llama «aprendizaje supervisado». Por el contrario, cuando el sistema analiza los patrones que pueda haber en los datos tal y como son, se llama «aprendizaje no supervisado». También hay un término medio. El «aprendizaje semisupervisado» se inicia con una semilla, o pequeña parte de datos, que ha sido preparada por los seres humanos, y a continuación la va proyectando cada vez sobre una mayor cantidad de datos sin supervisión.

En los últimos años, los científicos de IA se han centrado ampliamente en un

tipo de aprendizaje automático específico, el llamado «aprendizaje profundo», que ha ofrecido resultados impresionantes como enfoque de aprendizaje supervisado. A continuación comentaré el aprendizaje supervisado con cierto grado de detalle, además del aprendizaje profundo y sus aplicaciones. Como puedes suponer, el aprendizaje supervisado es un enorme manto que aglutina diferentes tipos de aprendizaje; los exploraré para ofrecer una visión general de los problemas que encuentra la IA.

Un tipo común de aprendizaje supervisado es la clasificación, que ha sido ampliamente tratada en centros de investigación y aplicaciones comerciales. Por ejemplo, los clasificadores aprendidos filtran el correo basura. El resultado de salida es un sí o no binario: el mensaje o bien es correo basura o no lo es. Por lo general, el sistema clasificatorio del correo basura se encuentra supervisado por el usuario de la cuenta de correo, que marca los mensajes entrantes como correo basura y los manda a la carpeta de correo no deseado o a la basura. En un segundo plano, el sistema de aprendizaje automático etiqueta aquellos mensajes que sean ejemplos positivos de correo basura. Cuando ya ha reunido una cantidad suficiente de ejemplos, el sistema aprende por sí mismo usando esos y otros mensajes entrantes, y crea un circuito de retroalimentación que converge en la diferencia entre los mensajes aceptables y los que son correo basura.

El filtro del correo basura es uno de los primeros ejemplos de la utilidad del aprendizaje automático en la red. Los algoritmos bayesianos ingenuos y otros clasificadores sencillos de probabilidad asignan puntuaciones numéricas a las palabras del mensaje indicando si son correo basura o no, y es el usuario quien facilita las categorías de correo deseado y no deseado. Al final, el clasificador acaba teniendo una hipótesis o modelo de correo basura basado tan solo en el análisis de las palabras de los mensajes. Los mensajes futuros se filtrarán de manera automática y los mensajes no deseados irán a parar a la carpeta de correo basura. Hoy en día, los clasificadores de correo basura utilizan muchísimo conocimiento suministrado por el hombre —pistas sobre lo que representa un correo no deseado, como ciertas palabras en el lema, términos y frases «basuriles» conocidos, etc—. Los sistemas no son perfectos, en buena medida por culpa del constante juego del gato y el ratón que se da entre los proveedores de servicios y los creadores de correo basura, que no dejan de probar enfoques nuevos y diferentes para engañar a los filtros entrenados.[3]

La detección del correo basura no constituye un ejemplo demasiado sexi de aprendizaje supervisado. Los sistemas de aprendizaje profundo modernos

también recurren a la clasificación para tareas como el reconocimiento de imágenes o de objetos visuales. Los populares concursos de ImageNet presentan ante sus participantes una tarea de aprendizaje supervisado a gran escala, con millones de imágenes que ImageNet ha descargado de sitios web como Flickr para usarlas como entrenamiento y poner a prueba la precisión de los sistemas de aprendizaje profundo. Todas esas imágenes han sido etiquetadas por personas (que han ofrecido sus servicios al proyecto a través de la interfaz Mechanical Turk de Amazon) y los términos que aplicaron han servido para establecer una base de datos estructurada de palabras inglesas conocida como WorldNet. Cada subconjunto seleccionado de las palabras de WorldNet representa una categoría que hay que aprender, usando nombres comunes (como «perro», «calabaza», «piano», «casa») y una selección de objetos más desconocidos (como «Scottish terrier», «mono rojo», «flamenco»). El concurso consiste en ver cuál de los clasificadores de aprendizaje profundo que compiten es capaz de etiquetar más imágenes de manera correcta, tal y como lo hicieron las personas. En los concursos de ImageNet se utilizan más de mil categorías, así que la tarea supera con creces el problema de sí/no que se presenta a los detectores de correo basura (o en cualquier otra labor de clasificación binaria, como la que se limita a etiquetar si la imagen pertenece a una cara humana o no). Participar en esta competición implica realizar una tarea masiva de clasificación usando los datos de píxel como estímulo.[4]

En las aplicaciones que procesan un lenguaje natural se suele usar la clasificación secuencial. Se trata las palabras como si tuvieran un orden definido, una secuencia. La clasificación documental o textual puede usar un enfoque más simple, sin ordenación —como un modelo «bolsa de palabras», BOW según su acrónimo inglés—, pero esa información adicional donde las palabras se ven como un texto ordenado suele mejorar el rendimiento de la clasificación de textos. Por ejemplo, las palabras que aparecen en el título y en el primer párrafo nos proporcionan a menudo pistas bastante potentes sobre el significado o el tema del artículo. La clasificación de textos puede explotar esos rasgos a la hora de autoetiquetar artículos con categorías como «CIENCIA», «NEGOCIOS», «POLÍTICA» y «DEPORTE». La clasificación de textos es otro ejemplo de aprendizaje supervisado, porque se inicia cuando los seres humanos etiquetan artículos debidamente con sus categorías y proporcionan un estímulo de entrada al sistema de aprendizaje. Igual que la colección de fotos correctamente etiquetadas de ImageNet, también hay corpus amplios, o conjuntos de datos creados por personas al comentar colecciones de textos, cuyos metadatos sobre temas y demás rasgos resultan útiles para la formación de los sistemas de

aprendizaje supervisado en labores de procesamiento del lenguaje.

El aprendizaje automático supervisado se encuentra tras una buena parte de la red contemporánea. Por ejemplo, posibilita la personalización de noticias y otras fuentes de contenido. Cuando el usuario hace clic principalmente sobre noticias de política, un algoritmo de aprendizaje supervisado que se está ejecutando en segundo plano (pongamos que en los servidores de Facebook) le presentará cada vez más noticias sobre política. Otros enfoques más sofisticados clasifican esas noticias políticas según su punto de vista; ofrecen noticias de corte más conservador o liberal al usuario cuya tendencia haya sido identificada por el sistema, e incluso las clasifican según sus sentimientos —como cuando un sistema clasifica los textos de opinión como positivos o negativos, igual que las críticas de cine.

Junto con la clasificación, los enfoques de aprendizaje supervisado se usan para etiquetar de manera automática los objetos individuales de una secuencia en vez de la secuencia entera, como sucede con la clasificación de imágenes y de textos. Es lo que se conoce como aprendizaje secuencial. Un ejemplo simple (aunque aburrido) de este es el etiquetado gramatical, donde una secuencia de palabras como «la vaca marrón» se etiqueta según sus funciones: «La / AD vaca / NC marrón / ADJ», refiriéndose esas etiquetas a sus caracteres respectivos de artículo determinado, nombre común y adjetivo. El aprendizaje secuencial no se sirve de las reglas de la lingüística para facilitar a los programas un conocimiento sobre las partes gramaticales del discurso; en cambio, las personas se limitan a etiquetar las palabras de las frases con su categoría gramatical correcta y proporcionan esos datos de preparación humana como entrada del algoritmo de aprendizaje. Las máquinas resolvieron hace mucho tiempo el problema del etiquetado gramatical; el suministro de decenas de miles de frases marcadas alcanza un rendimiento de nivel humano sobre los datos sin examinar, es decir, cualquier frase que no se haya utilizado durante el entrenamiento. Otro problema muy explorado dentro del procesamiento del lenguaje es el del reconocimiento de entidades, donde el sistema de aprendizaje supervisado predice las entidades que aparecen en un texto, como menciones a personas, lugares, momentos, empresas y productos. La frase «El señor Smith informó de que XYZ Co. vendió más de diez mil aparatos durante el primer trimestre» podría etiquetarse como «El señor Smith / PERSONA informó de que XYZ Co. / COMPAÑÍA vendió más de diez mil / NÚMERO aparatos /PRODUCTO durante el primer trimestre / FECHA».

La clasificación secuencial también se puede usar para realizar predicciones de series temporales, donde los objetos previos permiten predecir cuál será el objeto siguiente. Los sistemas de reconocimiento de voz como Siri son un tipo de predicción de series temporales, igual que los populares sistemas de voz a texto. La predicción de series temporales tiene importantes aplicaciones en tareas tan complejas como los diagnósticos médicos, la planificación industrial y el precio de las acciones, entre otros.

El aprendizaje supervisado es la causa de casi todos los éxitos de importancia que ha cosechado el aprendizaje automático hasta la fecha, incluyendo el reconocimiento de imágenes o de voz, la navegación autónoma de vehículos sin conductor y la clasificación de textos y estrategias de personalización en la red. El aprendizaje no supervisado presenta la virtud de requerir una preparación mucho menor de los datos, ya que no son los seres humanos quienes añaden las etiquetas a los datos de entrenamiento. Pero una consecuencia directa de esa pérdida de la «señal» humana es que los sistemas no supervisados llevan un gran retraso en comparación con sus primos supervisados a la hora de realizar las tareas del mundo real. El aprendizaje no supervisado resulta útil en tareas abiertas como la de permitir que el ser humano visualice los grandes conglomerados de datos que generan los algoritmos de aprendizaje no supervisado. Pero, ya que que la mayor parte del bombo relacionado con el aprendizaje automático —y en especial con el aprendizaje profundo— está relacionada con el aprendizaje supervisado, voy a centrar la discusión principalmente en él. Sin embargo, no olvidemos que todas las limitaciones de origen inductivo que presentan los enfoques de aprendizaje supervisado aparecen con mayor fuerza incluso en el aprendizaje no supervisado. Al centrarnos en el aprendizaje supervisado, estamos prestando atención al mejor y más potente de los casos.

EL APRENDIZAJE AUTOMÁTICO COMO SIMULACIÓN

Desde un punto de vista conceptual y matemático, el aprendizaje automático es intrínsecamente una simulación. Los diseñadores de cada sistema de aprendizaje automático examinan un problema de uso intensivo de datos, y si existe algún tratamiento posible de aprendizaje automático lo consideran «bien definido».

Asumen que alguna de sus funciones puede simular una conducta del mundo real o un sistema real. Se considera que el sistema real cuenta con un patrón oculto que da pie al resultado observable en los datos. La cuestión consiste en no recopilar de manera directa el patrón oculto —lo cual obligaría a comprender algo más que los datos—, sino en simular ese patrón oculto analizando las «huellas» que haya dejado en los datos. Es una distinción importante.

Piensa en otra tarea del procesamiento de lenguaje, la que se conoce como «etiquetado de roles semánticos». En ella, los diseñadores del algoritmo de aprendizaje desmontan el significado de las frases en términos de preguntas comunes del tipo quién, para quién, qué y cuándo. La función del algoritmo de aprendizaje consiste en tomar frases de ejemplo como entrada y generar el resultado de un conjunto de etiquetas que respondan a esas preguntas identificando los roles semánticos expresados en la frase. Por ejemplo, la frase puede contener a un agente que realiza una acción, un tema (el objeto involucrado en la acción) y un beneficiario de esta, y etiquetarse así: «John / AGENTE tiró /ACCIÓN la pelota / TEMA a Lizzy / BENEFICIARIO». En todos esos casos, el enfoque del aprendizaje automático implica asumir un comportamiento que se da pero que es desconocido, y utilizar un enfoque de aprendizaje para imitarlo de la mejor forma posible, descubriendo una función f. Como resultado del entrenamiento al que se somete al sistema se genera una f como modelo o teoría del comportamiento que se da en los datos. Este modelo puede regir roles semánticos, o entidades, o partes gramaticales, o imágenes de peces dorados; todo depende de la tarea de aprendizaje—. El aprendizaje automático es, por naturaleza, la simulación de un proceso que resulta demasiado complejo o que es desconocido, en el sentido de que no se dispone de unas reglas fáciles de programación, o de que entenderlo correctamente requeriría de un esfuerzo demasiado grande. A veces, un aprendizaje no supervisado revela la existencia de patrones en los datos. Pero son los seres humanos quienes identifican ese patrón después de un análisis; el algoritmo no sabe buscarlo. Y, si supiera, estaríamos hablando ya de un aprendizaje supervisado.

La mayoría de nosotros conocemos las funciones por las clases de matemáticas en el colegio, y su ejemplo más clásico es de tipo aritmético: 2 + 2 = 4 es una ecuación cuyo operador, el símbolo de la suma, es técnicamente una función. Las funciones generan respuestas únicas según su entrada: así, la función de suma nos devuelve un 4 para 2 + 2 (y nunca un 5, salvo en las novelas de George Orwell). Los primeros científicos de IA asumieron que se podrían resolver

numerosos problemas del mundo real suministrándoles reglas que equivalieran a funciones de salida conocida, como con las sumas. Sin embargo, resultó que la mayoría de los problemas de interés para los investigadores en IA tienen funciones desconocidas (si es que existe alguna función relacionada con ellos). Por ello, ahora tenemos el aprendizaje automático, que busca aproximarse o simular esas funciones desconocidas. Este carácter «falso» del aprendizaje automático pasa desapercibido cuando el rendimiento del sistema se acerca de manera notable al de los seres humanos, o lo mejora. Pero la naturaleza imitativa del aprendizaje automático se ve expuesta con rapidez cuando el mundo real se separa de la simulación aprendida.

Este hecho tiene una importancia enorme, y se pasa por alto demasiado a menudo en los debates sobre el aprendizaje automático. He aquí otro hecho: los límites del mundo de un sistema de aprendizaje automático quedan precisamente establecidos por los conjuntos de datos que se le proporcionen durante su entrenamiento. El mundo real no deja de generar conjuntos de datos: veinticuatro horas al día, siete días a la semana, a perpetuidad. Por ello, cualquier conjunto de datos dado es solo una fracción muy pequeña de tiempo que representa, en el mejor de los casos, una evidencia parcial del comportamiento de los sistemas del mundo real. Ese es uno de los motivos por los que la larga cola de acontecimientos improbables resulta tan problemática: el sistema no cuenta con una comprensión verdadera del sistema real (en comparación con el simulado). Esto es de una importancia tremenda para los debates sobre el aprendizaje profundo y la inteligencia artificial general, y plantea una serie de consideraciones problemáticas sobre cómo, cuándo y hasta qué punto deberíamos confiar en unos sistemas que técnicamente no comprenden los fenómenos que están analizando (salvo por lo expresado en sus conjuntos de datos durante el entrenamiento). Volveremos sobre estos temas en capítulos posteriores, ya que son capitales para comprender el paisaje del mito.

Existen al menos dos problemas con el aprendizaje automático como camino en potencia hacia la inteligencia general. Uno, que ya hemos tocado, es que se puede aprender de manera exitosa, al menos durante un tiempo, sin la menor comprensión del tema. Un sistema entrenado puede predecir resultados, entendiendo en apariencia el problema, hasta que un cambio o suceso inesperados hacen que la simulación se vuelva inútil. De hecho, las simulaciones que fracasan, cosa que sucede muy a menudo, pueden volverse peor que inútiles: pensemos en el uso del aprendizaje automático a la hora de conducir, y que contar con sus predicciones automatizadas provoque una falsa confianza. Es algo

que pasa por todas partes; el mundo real, tan desordenado, siempre está alterando su rumbo. Se cambia de tema de conversación. Las acciones siguen una tendencia al alza, pero acto seguido un suceso exógeno, como una reestructuración corporativa, un terremoto o una inestabilidad geopolítica, hace que se desplomen. Es posible que Joe se pirre por los blogueros conservadores hasta el día en que su amigo Lewis le sugiera una revista digital inclinada hacia la izquierda, que su fuente de noticias había descartado y le había ocultado por completo. Es posible que Mary adore los caballos hasta que Sally, el suyo, se muere y ella desarrolla una pasión por el zen. Etcétera. La verdad es que el nombre de «aprendizaje automático» resulta poco apropiado, ya que los sistemas no aprenden en el sentido en que lo hacemos nosotros, adquiriendo una valoración del sentido del mundo cada vez más sólida y profunda. Son más bien como curvas de campana del aprendizaje: simulaciones basadas únicamente en datos de aquello que experimentamos de una manera directa en el mundo real.

El sentido común sirve de mucho a la hora de comprender las limitaciones del aprendizaje automático: nos dice que la vida resulta impredecible. Por ello, la única crítica dañina de verdad que se le puede hacer al aprendizaje automático es que mire hacia atrás. Al basarse en observaciones procedentes de conjuntos de datos —esto es, observaciones previas— puede revelar patrones y tendencias que nos sean de utilidad. Pero todo aprendizaje automático es una fracción de tiempo procedente del pasado; ante un futuro abierto en el que los cambios son deseables, los sistemas han de pasar por un nuevo entrenamiento. El aprendizaje automático solo puede ir a la zaga del flujo de nuestra experiencia, simulando (lo que esperamos que sean) regularidades útiles. Es la mente —no la máquina— la que marca el camino.

EL APRENDIZAJE AUTOMÁTICO COMO UNA IA DÉBIL

La naturaleza imitativa del aprendizaje automático también ayuda a explicar por qué se encuentra permanentemente atascado en aplicaciones definidas de forma débil, y por qué sus progresos hacia una inteligencia artificial general son pequeños o inexistentes. Los problemas bien definidos que existen en el procesamiento de lenguaje natural, como la clasificación de textos, el etiquetado gramatical, el analizador sintáctico y el reconocimiento de correo basura, entre

muchos otros, deben ser analizados de manera individual. Esos sistemas tienen que ser objeto de un rediseño amplio, han de ser transferidos para que puedan solucionar otros problemas, aun cuando sean similares. Resulta irónico que a esos programas los llamemos «aprendices» porque, para nosotros, el significado del verbo «aprender» implica en esencia huir de los recursos limitados para alcanzar una comprensión más general de las cosas del mundo. Pero los sistemas que juegan al ajedrez no juegan al go, de mayor complejidad. Ni siquiera los sistemas de go juegan al ajedrez. Incluso el muy publicitado sistema Atari del DeepMind de Google generaliza solo entre diferentes juegos de Atari, y ni siquiera logró aprender a jugar a todos ellos. Los únicos que se le dieron bien fueron aquellos que seguían unos parámetros estrictos. Los sistemas de aprendizaje más potentes son mucho más frágiles y limitantes de lo que podríamos suponer. Pero tiene sentido, porque los sistemas no son más que simulaciones. ¿Qué otra cosa podíamos esperar?

Los problemas con la inducción que hemos comentado más arriba no se derivan de la experiencia per se, sino del esfuerzo por fundamentar el conocimiento y la inferencia solo en esa experiencia, que es precisamente lo que hacen los métodos de aprendizaje automático para llegar a la IA. No debería sorprendernos, por tanto, que esos métodos —y los que están centrados en datos— padezcan también todos los problemas de la inducción. Los datos solo son hechos que han sido observados y almacenados en ordenadores para poder acceder a ellos. Y la observación de hechos, por mucho que los analicemos, no nos conduce a una comprensión general ni a la inteligencia.

En esta verdad ya asumida sobre la investigación científica (y sobre la filosófica) ha surgido un giro moderno: la disponibilidad, más o menos reciente, de cantidades masivas de datos, que al menos en un principio debían fortalecer los sistemas de IA con unos «cerebros» y unas percepciones que antes no se encontraban disponibles. En cierto modo, eso es cierto, pero no en el sentido necesario para escapar a los problemas de la inducción. A continuación, nos fijaremos en el big data.

EL FIN DEL BIG DATA

El big data —o macrodatos— es una idea manifiestamente amorfa que en general se refiere al poder que tienen los conjuntos de datos de gran tamaño para posibilitar análisis y percepciones esenciales por parte de empresas y gobiernos (y de los investigadores en IA). El término apareció impreso por primera vez en un contexto científico en 1997: fue en un documento de la NASA que describía los desafíos que existían para visualizar datos usando la tecnología de gráficos informáticos del momento. Sin embargo, no prendió hasta que se volvió popular durante la década siguiente como un término multifunción en los ámbitos empresarial e informático. Al parecer, el concepto moderno de big data afloró en los debates sobre inteligencia empresarial, sobre todo en un informe del grupo Gartner fechado en 2001 y dedicado a los desafíos que presentaba ese campo. El informe destacaba «tres uves» —volumen, velocidad y variedad— para describir los rasgos de aquellos vastos conjuntos de datos que irían cobrando una importancia cada vez mayor a medida que los recursos informáticos se volvieran más potentes y baratos. Sin embargo, el informe no llegó a utilizar el término de big data.[5] En cualquier caso, el término sí comenzó a aparecer por todas partes a finales de la década de 2000, y en 2014 Forbes capturó el bombo y la confusión que lo acompañaban con un artículo titulado «12 definiciones de big data: ¿cuál es la tuya?»[6]

Quizá cueste un poco definirlos con precisión, pero el big data —órdenes de magnitud superiores de colecciones de datos— se encuentra en la vanguardia de la revolución informática en la ciencia y en la industria. En 2012, la administración Obama anunció una Iniciativa para la Investigación y el Desarrollo en Big Data con la que se pretendía «solucionar algunos de los desafíos más urgentes a los que se enfrenta el país».[7] Y al menos una empresa, la firma de análisis empresariales SAS, se apresuró a inventar un nuevo título ejecutivo: vicepresidente de big data. Bombo, sin duda, pero la excitación en torno a los macrodatos representaba también un reconocimiento de que, a menudo, una mayor cantidad de datos implicaba una mayor ventaja de cara a analizar los problemas en ordenadores cada vez más potentes.

No obstante, desde un principio se dio una confusión conceptual sobre la manera exacta en que los macrodatos «fortalecían» las percepciones y la inteligencia. Al principio se pensó que los macrodatos mismos eran responsables de la mejora de los resultados, pero, a medida que los métodos de aprendizaje automático fueron tomando vuelo, los investigadores comenzaron a otorgarles el mérito a los algoritmos. El aprendizaje profundo y demás aprendizajes automáticos y las técnicas estadísticas condujeron a mejoras evidentes. Sin embargo, el

rendimiento de los algoritmos quedó ligado a los grandes conjuntos de datos. En cualquier caso, la IA estaba demostrando mejoras en su rendimiento, y algunos problemas que hasta entonces no habían tenido solución la encontraron de repente con la entrada de más datos. Y fue esa expansión de la percepción —en los negocios y en la ciencia— lo que los investigadores y los expertos quisieron capturar. En palabras de Jonathan Stuart Ward y Adam Barker, científicos informáticos de la universidad de St. Andrews, «el big data está relacionado de manera intrínseca con el análisis de datos y con el descubrimiento de significados derivado de esos datos».[8] La IA llevaba décadas esforzándose por encontrar un significado en los datos; de repente, al añadirles más datos aún, ese significado parecía revelarse por todas partes.

En 2013, Viktor Mayer-Schönberger y Kenneth Cukier admitían ya en su best seller Big Data. La revolución de los datos masivos que «no existe una definición rigurosa de big data», pero sugerían de todos modos que los macrodatos son «la capacidad de la sociedad para emplear la información de una manera novedosa y generar percepciones útiles o bienes y servicios de gran valor», y que su llegada implicaba que en aquel momento había «cosas que se pueden hacer a gran escala, y no desde una escala menor, para extraer nuevas percepciones o crear nuevas formas de valor».[9] Señalaron historias exitosas en los sectores público y privado que no hubieran sido posibles sin aquel aumento de tamaño en los conjuntos de datos. Piensa, por ejemplo, en la empresa emergente Farecast, fundada en 2004 por Oren Etzioni, emprendedor y profesor de ciencias de la informática de la universidad de Washington, que Microsoft compró en 2008 por más de 110 millones de dólares. Etzioni, que en la actualidad dirige el Instituto Allen de Inteligencia Artificial en Seattle, usó macrodatos en forma de casi doscientos mil millones de registros sobre el precio de los vuelos para encontrar tendencias en sus picos y valles en función de los días previos a la salida. El rendimiento de Farecast subrayó la sensación de que el big data significaba nuevos acercamientos y competencias que emergían de aquella gran cantidad de números; partiendo del sistema de referencia de Etzioni, que solo usaba doce mil precios óptimos, el sistema no dejó de mejorar sus predicciones. Al alcanzar los miles de millones de precios óptimos de coste sobre el precio de los vuelos, ofrecía ya un gran valor para el cliente en forma de predicciones certeras sobre el momento en que había que comprar cada billete de avión.

El big data, en su momento un término de moda, representa ahora la nueva normalidad para los negocios impulsados por IA de todo el mundo. Walmart creó

Walmart Labs para aplicar las técnicas de macrodatos y de extracción de datos a sus desafíos logísticos —comprar, almacenar y enviar la mercadería de manera eficiente en respuesta a la demanda de los consumidores—. Amazon usaba los macrodatos antes de que se pusieran de moda, catalogando y haciendo el seguimiento de las compras en la red, y ahora los usa como datos con los que alimenta los algoritmos de aprendizaje automático que ofrecen recomendaciones de productos, búsquedas mejoradas y otras formas de personalización. Los macrodatos son una consecuencia inevitable de la ley de Moore: a medida que los ordenadores se vuelven más potentes, las técnicas estadísticas como el aprendizaje automático mejoran y aparecen nuevos modelos de negocio —y todo se debe a los datos y su análisis—. Lo que ahora denominamos ciencia de datos (o, cada vez más, IA) es en realidad una disciplina antigua a la que la ley de Moore y los volúmenes masivos de datos, facilitados en su mayoría por el crecimiento de la red, han dado nuevas alas.

Gobiernos y organizaciones sin fines de lucro no tardaron en sumarse y usar los macrodatos para predecirlo todo, desde la fluidez del tráfico hasta el porcentaje de reincidencia entre los presos con derecho a la libertad condicional. Mayer-Schönberger y Cukier cuentan que Nueva York contrató a unos expertos en big data de la universidad de Columbia para que realizaran un modelo predictivo sobre la probabilidad de explosiones de tapas de alcantarilla en la ciudad. (Solo en Manhattan hay más de cincuenta mil tapas de alcantarilla.) El proyecto fue un éxito y se ofreció como ejemplo de la manera en que aquellas nuevas percepciones y competencias resultaban posibles gracias al aumento de escala de los datos. Al fin y al cabo, los obreros humanos no pueden comprobar decenas de miles de tapas de alcantarilla cada día. También otros ámbitos, desde el procesamiento de fichas médicas hasta las iniciativas actuariales del gobierno con respecto al voto y las fuerzas del orden, ofrecen ejemplos que en apariencia apoyan la afirmación según la cual el tamaño y la calidad de los datos —el big data— han posibilitado esas nuevas percepciones y competencias.

El éxito de los macrodatos en la industria y otros sectores condujo con rapidez a declaraciones exageradas sobre el poder inferencial de los datos por sí solos. En 2008, las provocativas declaraciones de Chris Anderson, el director de Wired, para quien los macrodatos marcarían el fin de la teoría científica, representaron el punto culminante del sensacionalismo relacionado con el tema.[10] Los científicos y demás miembros de la intelligentsia se apresuraron a señalar que la teoría es necesaria, cuando menos porque un conjunto de datos no puede pensarse e interpretarse a sí mismo, pero el artículo se mantuvo como una

especie de expresión cultural sobre el éxito mareante del diluvio de los datos. En realidad, lo que ocurrió es que, al principio, un batiburrillo de viejas técnicas estadísticas en el uso de la ciencia de datos y el aprendizaje automático en la IA contribuyó a que las expectativas del big data emergente se vincularan erróneamente al volumen de datos mismo. Fue una proposición ridícula desde el primer momento; los datos son hechos y, una vez más, no pueden esclarecer nada por sí solos. Aunque esto solo se haya visto en retrospectiva, los éxitos tempranos del aprendizaje profundo con el reconocimiento de objetos visuales, en los concursos de ImageNet, señalaron el principio de una transferencia de entusiasmo entre el big data y los métodos de aprendizaje automático que se benefician de él; en otras palabras, hacia el nuevo y explosivo terreno de la IA.

De modo que el big data ha llegado a su cenit, y ahora parece estar desapareciendo del debate popular a la misma velocidad con la que llegaron a él. Tiene sentido que el foco recaiga sobre el aprendizaje profundo porque, al fin y al cabo, son los algoritmos, y no los datos por sí solos, los responsables de haber aplastado a los campeones humanos de go, los que dominan los juegos de Atari, los que conducen coches y todo lo demás. Y, de todos modos, el big data ha encontrado un nuevo hogar en la IA contemporánea, mientras que los enfoques basados en datos, como el aprendizaje automático, se benefician de los inmensos volúmenes disponibles de estos para entrenar los modelos y ponerlos a prueba. Tal y como comentaba un observador hace poco, los macrodatos se han convertido en IA big data.[11]

Ya está bien de big data. Pero aún tenemos pendiente el tema de la inferencia. Y, en particular, la manera en que los métodos basados en datos, como el aprendizaje automático, pueden superar la brecha entre las simulaciones superficiales basadas en datos y un conocimiento real, obtenido por una capacidad de inferencia más poderosa que la inducción. El problema inmediato es que el aprendizaje automático está inherentemente basado en datos. Ya lo había comentado más arriba, pero a continuación voy a volver sobre ello de manera más precisa.

LA RESTRICCIÓN EMPÍRICA

Los métodos basados en datos sufren en general lo que podríamos denominar como «una restricción empírica». Para entender esa restricción, deberíamos establecer un nuevo aspecto técnico del aprendizaje automático, conocido como «extracción de características». Al abordar un problema concreto, los científicos de IA suelen comenzar identificando los rasgos sintácticos, o marcas, en unos conjuntos de datos que ayudarán a los algoritmos de aprendizaje a concentrarse en obtener la salida deseada. La ingeniería de rasgos es en esencia una habilidad, y se paga muy bien a los ingenieros y especialistas que disponen de ella de cara a identificar rasgos de utilidad (también a los que tengan el talento para ajustar los parámetros del algoritmo, otro paso del entrenamiento exitoso). Una vez identificados, los rasgos se extraen de manera exclusivamente informática durante las fases de entrenamiento, de prueba y de producción. Y en ese carácter exclusivamente informático radica el quid de la cuestión. Los sistemas de aprendizaje profundo ofrecerían un rendimiento mucho mejor en la difícil tarea del reconocimiento de imágenes si tan solo pudiéramos dibujar una flecha sobre el objeto que deseamos identificar dentro de una foto repleta de objetos y de fondos —usando, pongamos, el software de Photoshop—. Pero, por desgracia, ese rasgo de origen humano no se puede añadir a otras fotos ni se puede preparar de esta manera, así que el rasgo no es extraíble sintácticamente y, por tanto, se vuelve inútil. Esa es la semilla del problema. E implica que los rasgos útiles para el aprendizaje automático siempre han de encontrarse entre los datos, y que los seres humanos no podemos proporcionar ninguna pista que la máquina no pueda explotar a su vez «en estado salvaje», al poner a prueba el sistema o después de lanzarlo para su uso.

La extracción de características se realiza en la primera fase, la de entrenamiento, y se vuelve a realizar cuando el modelo ya ha sido entrenado, en lo que se conoce como «la fase de producción». Durante la fase de entrenamiento, se proporcionan datos etiquetados en forma de entrada al logaritmo de aprendizaje. Por ejemplo, si el objetivo consiste en reconocer fotos de caballos, la entrada es la foto de un caballo, y la salida, la etiqueta CABALLO. El sistema de aprendizaje automático («aprendiz») recibe así fotos etiquetadas o marcadas de caballos en forma de parejas de entrada-salida, y la tarea de aprendizaje consiste en simular el etiquetado de imágenes de modo que solo aquellas que muestren un caballo reciban la etiqueta de CABALLO. Se prosigue con el entrenamiento hasta que el aprendizaje genera un modelo —que es un fragmento de código estadístico que representa la probabilidad de un caballo, dada la entrada— que responda (o no) a un criterio de precisión.

Llegado este punto, el modelo generado por el aprendiz se utiliza para etiquetar de manera automática nuevas imágenes, hasta ahora inéditas. Esta es la fase de producción. A menudo, un bucle de retroalimentación forma parte de esa producción, ya que un humano puede corregir las imágenes de caballos mal etiquetadas y devolvérselas al aprendiz para que continúe capacitándose. Esto puede seguir así de manera indefinida, aunque las mejoras en la precisión se irán volviendo cada vez de menor importancia. La interacción entre los usuarios de Facebook es un ejemplo de bucle de retroalimentación: cuando haces clic sobre un contenido, o etiquetas a un amigo en una foto, estás devolviendo datos al sistema de entrenamiento basado en el aprendizaje profundo de Facebook, que en todo momento estudia y modifica tu analítica de clics para seguir variando o personalizando futuros contenidos.

La restricción empírica es un problema para el aprendizaje automático porque no se puede usar toda la información adicional que desearías suministrarle al aprendiz. A diferencia de las labores de reconocimiento de imagen, que dependen de los datos de píxel como características, en la comprensión del lenguaje hay numerosos problemas que incorporan especificaciones adicionales, esto es, características identificadas por personas que han de ser extraídas por los sistemas cuando se entrenan y utilizan modelos.

Piensa en este problema sencillo del procesamiento de lenguaje, en el reconocimiento de entidades nombradas, donde un conjunto de etiquetas semánticas como PERSONA, ORGANIZACIÓN, PRODUCTO, LOCALIZACIÓN o FECHA son etiquetas de destino o salida, y la entrada es un texto de forma libre, quizá procedente de aportaciones de Facebook. Una empresa o individuo podría querer conocer todas las aportaciones que mencionan a cierta empresa —pongamos, Blue Box, Inc.—. La búsqueda por palabras clave de «Blue Box, Inc.», en la que coincidirían solo esas palabras, podría ignorar referencias más informales, como «Blue Box» o «blue box» o incluso «the box»,*b según el contexto. El objetivo del reconocimiento de entidades nombradas consiste en usar el aprendizaje automático sobre grandes cantidades de aportaciones etiquetadas de modo que esas menciones informales también sean identificadas correctamente como referencias a la empresa. Así, se da la necesidad de una extracción de características: un humano que etiquete todas las menciones a Blue Box, Inc. en una colección de aportaciones que se van a usar en un entrenamiento de datos y que se las mande al sistema, que generará un modelo para etiquetar las menciones a «Blue Box, Inc.» durante la fase de producción.

El sistema de Blue Box depende de las palabras de las aportaciones pero también de sus características: de la presencia en las aportaciones de menciones a empresas. Una vez más, las características son necesariamente sintácticas, porque deben ser extraídas de manera completamente automática durante la fase de producción. Esta es la restricción empírica clave. Las características pueden ser ortográficas, como que se compruebe la presencia de letras mayúsculas, o léxicas, como que se compruebe la presencia de las palabras «blue» y «box» en ese orden, y pueden incluir información como «que acaben en Inc. o Incorporated». Se puede ejecutar un etiquetador gramatical sobre los datos de entrenamiento para etiquetar categorías como nombres comunes y nombres propios —más características que se detectan sintácticamente—. Sin duda hay otras características posibles. La clave, de nuevo, consiste en que todas ellas, aunque en un primer momento hayan sido identificadas por personas, sean luego extraídas de manera puramente informatizada; de otro modo, el sistema no actuaría automáticamente durante la fase de producción autónoma.

Aquí viene el problema. Algunas señales de menciones a la compañía Blue Box, Inc. requerirán per se de una inferencia —por ejemplo, cuando entre los datos aparezcan pronombres y otras referencias—. Esto complica de manera inmediata la labor de aprendizaje. Si leo un estado de Facebook y veo que alguien habla de Blue Box, pero luego se refiere a ella en los comentarios diciendo algo así como «los beneficios de la empresa», la descripción «de la empresa» no es una característica admisible para el sistema de reconocimiento de entidades nombradas. Se ha de encargar de ello un subsistema de resolución de correferencia, lo cual introduce una tasa de error —hay que tener en cuenta que la correferencia es un problema mucho más difícil que el del reconocimiento de entidades nombradas—. Peor aún, quizá sepamos que Bob está hablando sobre Blue Box, Inc. en un debate sobre el precio de sus acciones, pero, puesto que no encontramos nada al respecto en los datos sometidos a análisis, no existe ninguna característica que pueda ser detectada por el sistema. Alguien podría contar una anécdota sobre el hecho de que el fundador de otra empresa, XYZ, Inc., «adoraba el color azul, y buscaba algo simple y memorable, así que decidió bautizar su sistema operativo como “Blue Box”». Por contexto, aquí «Blue Box» se refiere al producto, no a la compañía, pero el sistema de reconocimiento de entidades nombradas no puede utilizar esa información contextual durante su entrenamiento. ¿Por qué? Pues porque no puede extraerlo de la mera sintaxis, de su entrada durante la producción.

La restricción empírica forma parte integral del aprendizaje automático. Eso

quiere decir que durante la fase de entrenamiento solo se pueden usar características puramente sintácticas que los métodos automáticos descubran entre los datos. Un sistema en verdad inteligente necesita características o señales en un sentido más amplio, que no procedan tan solo de los datos procesados.

Aunque el reconocimiento de entidades nombradas sea una tarea más o menos simple dentro del procesamiento del lenguaje natural, incluso aquí podemos ver las limitaciones inherentes a los enfoques basados exclusivamente en datos. Una mención de Blue Box en una aportación acerca del producto se transforma con facilidad en un falso positivo, y queda etiquetada como si se refiriera a la compañía. Estos ejemplos podrían aparecer en la cola larga de apariciones improbables, pero son bastante comunes en el lenguaje cotidiano, y el aprendizaje automático, limitado por la restricción empírica, no puede abordarlos. Con todo esto pretendo decir que los datos por sí solos, macro o no, y los métodos inductivos como el aprendizaje automático presentan limitaciones inherentes que constituyen barricadas para el avance de la IA. Resulta que el de la inducción representa un problema importante de verdad para la IA moderna. Su ventana hacia el significado está ligada de manera directa a los datos, y eso limita y restringe el aprendizaje.

LA HIPÓTESIS DE FRECUENCIA

Además de la restricción empírica, los métodos de aprendizaje automático dependen de una desafortunada hipótesis de frecuencia. Como sucedía con la restricción empírica, vuelve a tratarse de una consecuencia directa del fundamento enumerativo de la inferencia inductiva —en realidad, se trata de una reafirmación—. Irónicamente, el valor de los macrodatos en el aprendizaje automático es en realidad una muestra de esa suposición: cuanto más, mejor. Los sistemas de aprendizaje automático son solo máquinas contadoras un poco más sofisticadas. Para seguir con el ejemplo de Blue Box, podríamos codificar una lista de características comprobando, pongamos, si esa palabra o secuencia de dos palabras de algunas aportaciones de Facebook aparece en un diccionario de palabras que incluya nombres de empresas como IBM, Microsoft, Blue Box, etc., o comprobando si le siguen Inc. o LLC, o si se trata de un acrónimo, o si la

primera letra está en mayúscula, o si es un nombre común.[12]

La suposición de frecuencia entra en juego porque, en general, a mayor frecuencia de resultados con esa característica, más útil resultará para el entrenamiento. En la ciencia de datos, esto es necesario; si las características de los datos son aleatorias, no se puede aprender nada (como comentamos antes). Pero si existe un patrón, primero, a causa de la restricción empírica, tiene que encontrarse en los datos; y segundo, como consecuencia, la única manera de determinar la intensidad de la asociación entre la entrada y la salida es a través de la frecuencia. ¿Cómo podría haber sido de otro modo? Si cada vez que «Inc.» sigue a un par de palabras la etiqueta en los datos de entrenamiento es EMPRESA, el aprendiz adjudica una probabilidad elevada a «Inc.» como característica de la salida deseada: empresa. Unos patrones que podrían pasar desapercibidos en miles de ejemplos cristalizan cuando los ejemplos son millones. Esa es la hipótesis de la frecuencia.

A las suposiciones de frecuencia se las puede poner patas arriba con la llamada detección de anomalías, que detecta transacciones bancarias fraudulentas o inicios de sesión indebidos. Esos sistemas también dependen de la hipótesis de frecuencia, ya que explotan lo que podríamos denominar la suposición de normalidad. Los sucesos normales hacen que los sucesos anormales resulten más llamativos. Si se pueden agrupar o concentrar miles o millones de inicios de sesión correctos por parte de los empleados, los raritos que quedan fuera del grupo llaman la atención. Por tanto, podrían ser ilegales o indebidos. Una vez más, el aprendizaje automático descubre lo que es normal —y, por tanto, lo que resulta anormal— analizando las frecuencias.

La suposición de frecuencia también sirve para explicar los «filtros burbuja» del contenido personalizado online. La persona que desprecia la política de orientación derechista acaba recibiendo solo opiniones de orientación izquierdista y otros contenidos informativos. El sistema basado en el aprendizaje profundo que controla esta salida no hace más que entrenar un modelo que, con el tiempo, pasa a reconocer el patrón de noticias que te gusta. Analiza tus clics y comienza a ofrecerte más de lo mismo. Las mismas observaciones pueden aplicarse a las sugerencias de Netflix, Spotify, Amazon y otros sitios web que ofrecen búsquedas personalizadas y una experiencia con recomendaciones. Esa conexión entre frecuencia de ejemplo (o características en ejemplos) y aprendizaje automático es intrínseca, esencialmente en el mismo sentido en que inferir que Todos los cisnes son blancos se vuelve cada vez más sencillo, porque

vas ganando confianza a medida que observas más y más cisnes blancos.

La suposición de frecuencia explica también la dificultad del problema de cola larga de los ejemplos anormales o inesperados. El sarcasmo, por ejemplo, resulta especialmente opaco para el aprendizaje automático, en parte porque es menos frecuente que el sentido literal. Resulta que el recuento funciona bien con ciertas tareas evidentes en la red, pero actúa en contra de las de corte más sutil. Si hay millones de ejemplos de ciudadanos enojados que tuitean «¡Trump es un idiota!», cualquier persona que tuitee «Trump es un idiota» con ánimo sarcástico y como respuesta ingeniosa a los críticos en su intento por superar en astucia a un rival, acabará reducido a una nueva instancia del patrón «Trump-idiota». Para comenzar, el algoritmo de aprendizaje no se encuentra en el negocio del conocimiento, así que para él el ejemplo es solo otra secuencia de palabras. El sarcasmo no es una característica basada en palabras, y tampoco se presenta con la frecuencia del sentido literal. El aprendizaje automático se muestra notablemente obtuso ante esos fenómenos del lenguaje —para disgusto de empresas como Google—. Le encantaría detectar el sarcasmo a la hora de dirigir los anuncios. Por ejemplo, si «¡Tráete la crema solar!» es un comentario sarcástico en una entrada sobre una tormenta de nieve, un sistema de colocación de anuncios sensible al contexto intentaría emitir en su lugar el anuncio de unos calcetines calefactables a pilas.

La hipótesis de frecuencia se vuelve aún más pronunciada cuando la entrada son artículos enteros de prensa que, pongamos, incluyan un texto con la tarea de clasificación mencionada anteriormente. Las noticias «curiosas» o «raras» que abundan en la red como lecturas ligeras representan una pesadilla para el aprendizaje automático, porque el significado de la palabra no es literal. Por ejemplo, un sistema de aprendizaje automático podría clasificar historias que describan sucesos «tontos» pero que técnicamente incluyan referencias a crímenes o delitos como casos genuinos de historias criminales. ¿Por qué no habría de hacerlo? Las noticias peculiares son menos frecuentes que aquellas otras que proporcionan información de manera directa, así que también aparecen menos en los conjuntos de entrenamiento —y, de todos modos, detectar el motivo por el que resulta peculiar aboca a otro problema con la restricción empírica—. Una noticia se reconoce como tonta o sarcástica cuando alguien comprende la voluntad del autor, lo que pretende comunicar. Pero las palabras en sí de esa historia pueden, por sus frecuencias en los datos de entrenamiento, apuntar a categorías bien definidas como las de política, deportes, crimen, etc. La pieza no podrá ser clasificada, ni comprendida correctamente, a menos que

los fragmentos sintácticos que la constituyen —las palabras— se interpreten desde una ventana de significado mucho más amplia. A falta de esta capacidad no inductiva, el sistema de aprendizaje automático pone en valor por defecto las frecuencias, y yerra el tiro. He aquí, por ejemplo, una noticia de Associated Press que apareció publicada una vez en Yahoo! News:

¡LOS TACOS O LA VIDA!

Fontana, California — El hambre de carnitas estuvo a punto de provocar una carnicería cuando a un hombre de Fontana le robaron un paquete de tacos a punta de pistola.

El sargento de policía Jeff Decker dijo que la víctima, de 35 años de edad, acababa de comprar tacos por valor de veinte dólares en un puesto callejero el domingo por la noche y regresaba en bicicleta a su casa cuando el sospechoso se plantó ante él y le dijo: «Dame los tacos».

Según Decker, el sospechoso cogió la bolsa con la comida, pegó un puñetazo en la cara a la víctima y se dispuso a huir.

Cuando la víctima le exigió que le devolviera los tacos, el sospechoso le apuntó con lo que pareció ser una pistola y amenazó con matarle antes de escapar.

Un sistema de clasificación de textos identificaría con facilidad esta noticia como una historia criminal: «sospechoso, víctima, huir, pistola». No obstante, la mayoría de los lectores humanos tendrán la impresión de que se trata de una historia cómica —o por lo menos no la vemos como el ejemplo típico de una historia de crímenes—. Se informa sobre los actos criminales porque son graves y preocupantes, pero una frase inicial del tipo «El hambre de carnitas...» señala que la voluntad de Associated Press consiste en informar sobre la historia de manera humorística. Hasta un niño de primaria se dará cuenta de esa voluntad, pero los sistemas de IA estarán encantados de clasificar el artículo como una nueva historia criminal ambientada en Fontana, California. La frecuencia es la asesina del humor. Cuenta el número de artículos periodísticos que incluyen víctimas, pistolas, amenazas y sospechosos que se dan a la fuga. Son de crímenes. El problema del supuesto de frecuencia ante esos ejemplos es que no

se dispone de ninguna solución conocida al usar el aprendizaje automático. El sentido de la historia se pierde dado su método, que analiza las palabras sintácticamente y cuenta la frecuencia de las palabras como señal de una categoría. Incluso con ejemplos más o menos sencillos, como el que nos ocupa, esa senda nos conduce a un callejón sin salida respecto a la inteligencia artificial general.

He aquí otra historia de Associated Press que recogieron diversos periódicos:

Un niño de once años muerde a un pit bull para defenderse de su ataque

São Paulo, Brasil — Un niño de once años se ha convertido en el centro de atención de la prensa brasileña después de haber clavado los dientes en el cuello del perro que le había atacado.

Los periódicos locales informaron el jueves de que Gabriel Almeida estaba jugando en el patio de la casa de su tío, en la ciudad de Belo Horizonte, cuando una pit bull llamada Tita se lanzó sobre él y le mordió en el brazo izquierdo. Almeida sujetó al perro por el cuello y le devolvió el mordisco... con tanta fuerza que se le rompió un colmillo.

En declaraciones al periódico O Globo, Almeida afirmó: «Es mejor perder un diente que la vida».

Unos albañiles que trabajaban cerca del lugar ahuyentaron al perro antes de que pudiera atacar de nuevo.

Sin duda, la noticia cuenta con un lado serio, pero desde luego que no se trata de una historia sobre el ataque de un pit bull. Tampoco se trata de la historia sobre un concurso de mordiscos entre un niño brasileño y un perro. Puesto que el niño no sufrió heridas de gravedad, pese a que perdió un diente al morder al perro, queda claro que el motivo para publicar la pieza no fue el de informar sobre el ataque de un can brasileño, sino el de subrayar lo extraño o divertido de ese contraataque sorpresa. Su contenido improbable —niño muerde a perro— es lo que la vuelve noticiable. En un caso como este, la IA y el aprendizaje absoluto no nos ayudan en nada. Nos hacen daño. Yerran el tiro por completo. Los

sistemas de aparente inteligencia artificial general que usen solo el aprendizaje automático serán, en el mejor de los casos, unos eruditos idiotas y molestos.

En esencia, la teoría subyacente de la inferencia se encuentra en el meollo del problema. La inducción requiere que la inteligencia sea resultado del análisis de datos, pero la inteligencia llega al análisis de los datos como paso previo y necesario. Siempre cabe esperar que los avances en la extracción de características o el diseño de algoritmos conduzcan en el futuro a una teoría más completa de la inferencia informática, pero deberíamos mostrarnos muy escépticos a ese respecto. Son precisamente la restricción empírica y la hipótesis de frecuencia las que limitan el alcance y la efectividad de las características detectables —que, al fin y al cabo, se encuentran en los datos para ser analizadas sintácticamente—. Es otra manera de decir lo que los filósofos y científicos de todos los colores aprendieron hace mucho tiempo: no basta con la inducción.

LA SATURACIÓN DE LOS MODELOS

El aprendizaje automático y el big data presentan otro problema, conocido como «saturación», que afecta la esperanza de obtener una inteligencia artificial general. La saturación se da cuando, al añadirle más datos —más ejemplos— a un algoritmo de aprendizaje (o técnica estadística), no se suma nada al rendimiento de los sistemas. No existe ningún entrenamiento que pueda prolongarse para siempre e ir ofreciendo una precisión cada vez mayor para un problema. Con el tiempo, el añadido de datos deja de incrementar el rendimiento. Los sistemas exitosos alcanzan una precisión aceptable antes de saturarse; si no es así, el problema no se podrá resolver utilizando el aprendizaje automático. El saturado es un modelo definitivo, que no mejorará por mucho que se le añadan más datos. En algunos casos podría incluso empeorar, aunque las razones para ello son demasiado técnicas para explicarlas aquí.

Rara vez se habla de la saturación del modelo, sobre todo porque numerosos problemas recientes no han dejado de beneficiarse del aumento de datos preparados. Pero los investigadores saben que la saturación resulta inevitable, y que acaba limitando el rendimiento de los sistemas de aprendizaje automático. Allá por 2013, Peter Norvig, director de investigación en Google, confesó a The

Atlantic la ansiedad que le generaba la saturación: «Podríamos dibujar la siguiente curva: a medida que vamos ganando en datos, ¿hasta qué punto mejora nuestro sistema? —se preguntaba—. Y la respuesta es que continúa mejorando..., pero estamos llegando a un punto en el que obtenemos menos beneficios que en el pasado».[13]

En el momento de escribir estas líneas, el comentario admonitorio de Norvig tiene siete años. Lo más probable es que los concursos de ImageNet no puedan usar más datos —los mejores sistemas disponen ya del 98 % de precisión (usando la medida de prueba estándar de obtener una etiqueta de destino entre las cinco primeras predicciones de un sistema)—. Pero los coches sin conductor, que creíamos a la vuelta de la esquina, continúan en fase de investigación intensa, y sin duda parte del problema radica en los datos de entrenamiento de las fuentes de vídeo etiquetadas, que no resultan insuficientes en número, pero sí inadecuadas para afrontar problemas de cola larga con escenarios de conducción atípicos que no obstante se han de tener en cuenta por razones de seguridad. Los modelos se están saturando, tal y como predijo Norvig. Es evidente que existe la necesidad de nuevos enfoques. Esas consideraciones son uno de los motivos por los que el llamado crecimiento en escala desde un éxito inicial hasta otro en toda regla resulta ingenuo y simplista. No se puede hacer que un sistema crezca en escala de manera indefinida. El aprendizaje automático —el aprendizaje profundo— no es una fórmula milagrosa.

En sus textos de 1950, Turing se declaró esperanzado ante la idea de que los sistemas computacionales pudieran aprender aquello que desconocían. El aprendizaje automático no era por entonces un término de IA, aunque ya se conocía la posibilidad de crear redes neuronales simples. Pero lo que Turing tenía en mente era una idea ampliada del aprendizaje, más parecida a la versión humana de este. No se podía programar las máquinas con todo el conocimiento que necesitarían, así que debía darse algún tipo de aprendizaje. Pensó que este podría surgir con la inducción. Algunas proposiciones, reflexionó, «vendrían “proporcionadas por la autoridad”, pero otras debería generarlas la máquina misma, por ejemplo, a través de la inducción científica». Desde su localización, en el ecuador del siglo XX, Turing había abandonado sus preocupaciones sobre la necesidad de percepciones externas a los sistemas formales. O, mejor dicho, tenía la esperanza de que encontraran su lugar en la nueva tecnología computacional.

Sin embargo, ni siquiera los científicos mismos usan la «inducción científica» en

el sentido que quiso darle Turing. Realizan conjeturas, acto seguido las ponen a prueba y acto seguido realizan más conjeturas. Turing nunca mencionó el trabajo de Peirce sobre la inferencia lógica. Al parecer, él no tenía un conocimiento sustancial de la inferencia abductiva en el sentido que le dio Peirce.

Y nosotros seguimos buscando sus máquinas capaces de aprender.

Capítulo 12

La inferencia abductiva

El origen de la inferencia como especulación

Charles Sanders Peirce estaba trabajando para el Servicio de Costas de Estados Unidos en un problema científico de cierta importancia. Peirce balanceaba péndulos. Péndulos delicados. El Servicio de Costas los utilizaba para medir variaciones en la gravedad de la Tierra, como parte de la ciencia gravimétrica. Esta disciplina integrada en la geodesia, que en el siglo XIX seguía siendo un ámbito en desarrollo, contribuía al estudio de la forma y el tamaño de nuestro planeta. Se necesitaban mediciones precisas de la topografía terráquea para todo, desde el proceso de excavación para construir nuevos edificios de oficinas hasta para ir a la guerra. Ese era el trabajo de Peirce.

La gravimétrica también requiere que se mida con precisión el tiempo. Y así fue que, en el verano de 1879, Peirce se encontró a bordo de un vapor costero que zarpó de Boston con destino a Nueva York, en posesión de un reloj muy caro que debía usar en su trabajo con los péndulos. Él mismo lo había elegido, y le costó trescientos cincuenta dólares, una cantidad enorme en aquel momento. El Servicio de Costas había pagado la factura. Por la mañana había desaparecido, igual que el resto de sus posesiones.

Peirce era famoso por perder y extraviar equipos caros por todas partes del mundo, allí donde el Servicio de Costas lo hubiera enviado a realizar sus mediciones gravitacionales. El robo del reloj encajaba dentro de un patrón que, poco a poco, había ido provocando fricciones entre Peirce y el gobierno norteamericano. Por ello, la sustracción del reloj y su labor detectivesca para recuperarlo tuvieron una gran importancia personal para él. No es de extrañar que más tarde se tomara todo aquello como una gran clase de inferencia.

Las cosas fueron así, según su propio recuento: Peirce recordó que, después de que el barco atracara en el puerto de Nueva York por la mañana, cogió un taxi para asistir a un congreso en la ciudad. Al llegar, se dio cuenta de que se había

olvidado el reloj, con su fina cadena de oro, y su abrigo bueno, y se apresuró a volver para recogerlos. Al entrar en el camarote vio que ambos habían desaparecido, y comprobó que no se debía a que el capitán se los estuviera guardando. Se trataba de un robo, era evidente, y tenía que haberlo cometido uno de los asistentes de camarote del barco.

Con la ayuda del capitán, Peirce pudo reunir a todos los asistentes en la cubierta de la embarcación. Los colocó en fila y comenzó a dirigirles una perorata mientras les iba mirando a la cara, uno por uno. Esperaba detectar alguna indicación clara de culpabilidad, pero se llevó una decepción: «No tengo ni el más mínimo ápice de luz para proseguir», admitió para sí. Pero, cuando comenzaba a alejarse, pensó: «Debes señalar a alguien. No importa que no tengas motivos, has de decir quién crees que es el ladrón».

Peirce regresó sobre sus pasos para mirar al grupo de nuevo... y, de repente, «cualquier sombra de duda se desvaneció».[1] ¿Qué había pasado? Pues que había estado elaborando los detalles de sus inferencias «conjeturales»: la abducción. Ahí iba una suposición. Un ejemplo de la vida real —si en efecto tenía razón—. Peirce se volvió hacia el hombre al que había señalado como el culpable y, después de pedirle que abandonara la fila y le acompañara al camarote, le ofreció un trato: cincuenta dólares a cambio de que devolviera los objetos robados.

«Bien —dijo—, el billete es tuyo si te lo trabajas. No quiero averiguar quién me ha robado el reloj, si puedo evitarlo, porque si lo hiciera estaría obligado a mandarle a Sing Sing [la prisión de Nueva York], lo cual me costaría más de cincuenta dólares. Y además lo sentiría en el alma por ese pobre idiota que se ha creído mucho más listo que los hombres honestos.»[2]

Resulta que el acusado no sabía gran cosa sobre la inferencia abductiva, o quizá pensó que estaba viéndole el farol a Peirce. «Pues vaya, sí me gustaría ganarme esos cincuenta dólares —contestó—, pero resulta que no sé nada sobre sus cosas. Así que no puedo.»[3]

Incapaz de extraerle una confesión, Peirce acudió con presteza a la agencia de detectives Pinkerton (un lugar «formidable»), y se reunió con el director de la oficina de Nueva York, George Bangs, para pedirle que siguieran al hombre, ya que estaba seguro de que pensaba llevar el reloj a una casa de empeños, donde le darían unos cincuenta dólares por él. Bangs desdeñó las conjeturas de Peirce, prefirió utilizar las reglas habituales y los métodos ya conocidos para

concentrarse en los sospechosos. Mandó a un detective a que comprobara a los asistentes de camarote y se descubrió que uno de ellos tenía antecedentes criminales con cargos que incluían el carterismo y otros delitos menores; era el sospechoso más probable. Sin embargo, la vigilancia a la que le sometieron los detectives no arrojó ningún resultado. Peirce, que seguía convencido de que el sospechoso original habría acudido a una casa de empeños, aceptó el consejo de Bangs y ofreció una recompensa considerable, de ciento cincuenta dólares, por la información que le permitiera recuperar sus posesiones.

El anuncio que publicó tuvo efecto en menos de un día; un prestamista se presentó con el reloj. Y, al oír la descripción del hombre que se lo había vendido, Peirce obtuvo la confirmación que buscaba: era la viva imagen del asistente al que él había acusado.

Peirce se hizo con su dirección (Bangs se estaría divirtiendo) y se presentó sin avisar en el apartamento del hombre. Le abrieron la puerta dos mujeres, que de inmediato amenazaron con llamar a la policía. Peirce no les hizo caso, sino que se dirigió hacia un amplio baúl de madera que había entrevisto. En el fondo encontró la cadena de su reloj de oro, junto con la bitácora y la brújula que iban sujetos a ella.

Mientras tanto, una de las mujeres había escapado por el apartamento contiguo. Tal y como relató Peirce, después de llamar a esa puerta y de que le abrieran dos chicas jóvenes, entrevió un fardo doblado con pulcritud encima de un piano y «abriéndome paso con suavidad entre ellas» se dirigió a recuperar el abrigo.

Es posible que Peirce se imbuyese de Dupin. Juró que lo había conjeturado todo. El señor Bangs se encogió de hombros.

Peirce concluye el artículo en el que cuenta la historia, «Conjeturas», señalando que, a primera vista, estas pueden ser motivo de recelo: «Supongo que casi todo el mundo habrá tenido experiencias similares».[4]

Sí que conjeturamos. Y nuestras conjeturas —inferencias— nunca son incuestionables. Pero el misterio es: ¿por qué nuestras corazonadas, nuestras conjeturas, no son más que aproximaciones aleatorias a la verdad?

Al principio de «Conjeturas», Peirce se pregunta cómo lograron «Galileo y los demás maestros de la ciencia» alcanzar sus teorías verdaderas tras tan escasas suposiciones fallidas. Los científicos, como todos nosotros, deducen sus

explicaciones a partir de lo que saben y observan. Queremos subsumir esas inferencias en nuestro flujo de observaciones, en los hechos. Pero gran parte de lo que inferimos escapa al marco de la observación pura. El conocimiento contextual impregna prácticamente todas las inferencias que realizamos. Que Peirce se sirva de Galileo para apuntalar su historia resulta, por tanto, adecuado: a menudo, los descubrimientos científicos se atribuyen al seguimiento meticuloso de unos métodos ya conocidos, pero eso no es del todo cierto. Escondemos el misterio detrás del método. Galileo también conjeturó, igual que Peirce a bordo del vapor. En ambos casos, la investigación subsiguiente demostró que, de algún modo, aquella conjetura no iba desencaminada.

Peirce comparó la conjetura con un instinto que permite seleccionar, entre «al menos mil millones» de hipótesis posibles, la que parece correcta. Cuando Holmes conoce a Watson, le pregunta si acaba de volver de la guerra, porque repara en su bronceado y en su cojera. Un médico militar que ha regresado de la guerra en Afganistán, supone. ¿Se trata solo de una conjetura? No: es una inferencia.

Cuando intentamos comprender hechos particulares —como el robo de un reloj — en vez de circunstancias regulares, nos vemos abocados de manera inevitable a una especie de juego de manos, la selección o invención de una hipótesis que pueda explicar el hecho. La inducción pasa de los hechos a las generalizaciones que nos otorgan el conocimiento (nunca basado en certezas) de algo regular. Pero la abducción pasa de la observación de un hecho en particular a una regla o hipótesis que lo explique. La abducción se encuentra íntimamente ligada al razonamiento que va del hecho a su causa —en el ejemplo de Peirce, desde el hecho del robo a su causa, el ladrón—. Sherlock Holmes se refería a ese tipo de razonamiento como «nada más que sentido común», y en buena medida tenía razón. Pero el sentido común resulta en sí mismo misterioso, precisamente porque no encaja en un marco lógico, a diferencia de la deducción o de la inducción. La abducción capta la percepción de que buena parte de nuestro razonamiento diario es una especie de labor detectivesca, en la que vemos los hechos (datos) como pistas que nos ayudan a encontrar el sentido de las cosas. Hacer hipótesis es algo que se nos da extraordinariamente bien, lo cual, en opinión de Peirce, no se puede explicar desde la mecánica, sino a través de una operación mental a la que él llama, a falta de otra explicación, instinto. Conjeturamos qué hipótesis parecen probables o verosímiles dentro de un contexto de posibilidades en efecto infinitas.

Es algo que debemos tener en cuenta a la hora de construir una inteligencia, porque se trata del punto de partida de cualquier pensamiento inteligente. Sin un paso abductivo previo, las inducciones son ciegas y las deducciones se tornan inútiles también.

La inducción requiere el primer paso de la abducción porque necesitamos poner en observación una infraestructura que dé sentido a los que los filósofos llaman los datos de los sentidos —la experiencia cruda, sin interpretar—. Incluso en una inducción simple, donde inducimos la proposición general de que «Todos los cisnes son blancos» a partir de la observación de esos cisnes, una infraestructura conceptual mínima, o teoría, guía la adquisición de conocimiento. Podríamos inducir que todos los cisnes tienen pico siguiendo la misma estrategia inductiva, pero esa inducción tendría menos fuerza porque todas las aves tienen pico y los cisnes forman parte de un pequeño subconjunto de las aves. Usamos el conocimiento previo para generar hipótesis. La intuición proporciona problemas de interés a los matemáticos.

Cuando los desarrolladores de DeepMind aseguraron, en un artículo muy leído de la prestigiosa revista Nature, que este había llegado a dominar el go «sin conocimientos humanos», malinterpretaron la naturaleza de la inferencia, mecánica o de otro tipo. Es evidente que el artículo «exageró el caso», en palabras de Marcus y Davis.[5] De hecho, los científicos de DeepMind diseñaron un modelo admirable del go en AlphaGo, y se tomaron la molestia de buscar los mejores algoritmos para solucionar varios aspectos del juego —todo ello antes de que el sistema llegara a participar en una competición real—. Tal y como explican Marcus y Davis, «el sistema era muy dependiente de cosas que los investigadores humanos habían descubierto durante las décadas anteriores para lograr que las máquinas participaran en juegos como el go, siendo la más importante el árbol de búsqueda Monte Carlo ... el muestreo aleatorio a partir de un árbol de diferentes posibilidades de partidas, lo cual no tiene nada que ver, en un sentido intrínseco, con el aprendizaje profundo. Además (a diferencia [del sistema de Atari]), DeepMind llevaba incorporadas las reglas y otros conocimientos detallados del juego. La información según la cual no tenía relación con el conocimiento humano no se ajustaba a los hechos».[6] Una forma más sucinta de expresarlo consiste en decir que el equipo de DeepMind usó inferencias humanas —en concreto, las de tipo abductivo— para diseñar el sistema y que este completara su tarea con éxito. Esas inferencias fueron suministradas desde fuera del armazón inductivo.

¡SORPRESA!

Peirce comprendió que, en su origen, la abducción es una reacción a la sorpresa:

El hecho sorprendente C es observado.

Pero, si A fuera verdad, C sería su consecuencia natural.

Por consiguiente, existen motivos para sospechar que A es verdad.[7]

Las sorpresas aparecen en la larga cola de problemas de la inducción. Y las inferencias abductivas buscan la explicación de hechos particulares (A), no leyes ni generalizaciones, como la inducción. C también es un hecho particular —y sorprendente—. Así que la abducción no es en absoluto una generalización.

Las inferencias que van de las observaciones particulares a las explicaciones particulares forman parte de una inteligencia normal. Si la camarera Kate trabaja por lo general en un Starbucks los jueves, pero no los viernes, un ordenador que hubiera cosechado su conocimiento a través de experiencias previas no esperaría su presencia un viernes, pero se enfrentaría a un problema de cola larga si al fin y al cabo se la encuentra trabajando ese día. Es posible que Kate esté haciendo horas extra, o que la hayan llamado ese día para que cubra una baja por enfermedad. Y quizá no trabaje el jueves, porque esté enferma o porque la hayan trasladado a otra tienda o porque haya dejado el trabajo. Todos estos son hechos particulares (sorprendentes) que pueden explicar su presencia o su ausencia. Se trata de inferencias de sentido común que no dependen de generalizaciones ni de expectativas. (Las investigaciones criminales, por cierto, siempre se inician con hechos sorprendentes. La inducción puede indicarnos que los hombres jóvenes cometen la mayoría de los crímenes, pero el investigador sigue necesitando saber quién ha sido el responsable del caso particular que le ocupa... y el culpable podría no ser ni hombre ni joven, como vimos en la calle Morgue.)

Peirce entendió la abducción como una forma debilitada de inferencia, en el sentido de que era conjetural —una abducción a la hora t podría revelarse errónea a la hora t + 1—. Numerosas inferencias en el mundo real se vuelven revisables; es decir, que se tornan erróneas o incompletas a raíz de observaciones subsiguientes o del aprendizaje (leyendo un libro, pongamos).

La inferencia conjetural es una característica, no un error, de los sistemas inteligentes. El robot Rosie puede creer que Kate ha dejado Starbucks porque un compañero de trabajo le ha proporcionado esa información, pero cuando Kate se presente dispuesta a trabajar diez minutos más tarde y el compañero sonría, el robot Rosie tendría que retractarse de esa inferencia. A duras penas reparamos en la velocidad a la que conjeturamos razones verosímiles para aquello que vemos (o sobre lo que leemos), y también en la rapidez con la que abandonamos o actualizamos esas conjeturas. El mundo cotidiano es un flujo constante de hechos en apariencia sorprendentes sobre un trasfondo de expectativas. Como los semáforos, buena parte del mundo no es una sorpresa constante..., pero es que los semáforos también se rompen.

El significado de la observación misma también atraviesa una alteración conceptual con la abducción. Mientras que la inducción trata la observación a modo de hechos (datos) analizables, la abducción ve el hecho observado como una señal que apunta hacia una característica del mundo. Esas señales pueden verse como pistas, porque se entiende desde un primer momento que están incrustadas en una red de posibilidades que podría apuntar hacia —o arrojar luz sobre— un problema particular o un asunto importante para el observador. En contextos culturalmente ricos como el de la resolución de crímenes, las pistas resultan necesarias porque hay que analizar demasiados datos y solo unos pocos son relevantes. En efecto, el problema básico que se deriva de usar métodos conocidos en el trabajo detectivesco es que los crímenes de difícil resolución o que en apariencia no la tienen tampoco encajan con las regularidades, y la acumulación de hechos no apunta en ningún sentido. Lo que buscan los detectives listos son pistas.

Igual que los cazadores. Estos no son solo observadores astutos, sino que observan cierto tipo de cosas. Huellas, excrementos, mechones de pelo, ramas rotas, aromas... Todo ello son pistas sobre la localización de la presa. Como los detectives en la escena de un crimen, los cazadores se dan a una búsqueda deliberada de pruebas sobre un acto reciente; observan resultados procedentes del pasado.

De manera quizá contraria a la lógica, las pistas no se consideran únicas. El cazador que se cruce con un aroma desconocido asumirá que se dirige hacia algo interesante, pero no completamente singular, porque en ese caso el aroma podría no funcionar como pista. El cazador que razone que un aroma desconocido puede pertenecer a una especie hasta el momento desconocida o a un extraterrestre no avanzará en su interés por encontrar a la presa.

Así, el cazador se interesa por una conjetura que encaja en su propósito específico. Es posible que el aroma sea único porque viene de un animal en celo. Por tanto, el aroma es un hecho sorprendente que puede explicarse conjeturando su origen a partir de los cambios experimentados en el animal durante la época de celo. Durante ese período, un animal conocido exuda aromas diferentes. Tengamos en cuenta que el cazador no dispone de un conocimiento previo de ese fenómeno (de ahí la inferencia abductiva), pero su razonamiento tiene lugar dentro de una estructura que excluye las posibilidades lógicas que no le permitan avanzar hacia el objetivo. Sorprendentemente, es probable que esas conjeturas acaben siendo ciertas —y ese es el motivo por el que Peirce le dio vueltas al «instinto conjetural» como meollo de tantas ideas inteligentes.

La visión es un ejemplo tan bueno como el de la caza. Incluso juzgar si un objeto concreto es una azalea, tal y como señala Peirce, implica una percepción física dentro de una red profunda de conocimientos previos y expectativas. Recuerda sus palabras:

Al mirar por la ventana en esta hermosa mañana de primavera veo una azalea en plena floración. ¡No, no! No es eso lo que he visto, aunque sí se trate de la única manera en que puedo describirlo. Es una proposición, una frase, un dato; pero lo que percibo no es una proposición, una frase, un dato, sino apenas una imagen que yo hago inteligible en parte a través de la exposición de un hecho. Esa exposición es abstracta, pero lo que yo veo es concreto.[8]

Podría parecer que los éxitos recientes en el uso de redes neuronales convolucionales (aprendizaje profundo) sobre tareas de reconocimiento visual de objetos, con los concursos de ImageNet como ejemplo rotundo, contradicen la insistencia de Peirce en que la inferencia abductiva sustenta la capacidad

perceptiva visual incluso cuando esta es en apariencia trivial. Sin embargo, esos éxitos aparentes prueban de hecho la idea de Peirce, ya que hay que reconocerle a la comunidad investigadora que haya señalado la fragilidad de esos sistemas en una bibliografía cada vez más amplia que aborda no solo la cuestión central de la inferencia, sino también las dudas acerca de su confianza y fiabilidad, así como su potencial para un uso erróneo. En palabras de la científica informática Melanie Mitchell, es ridícula la facilidad con que se puede llegar a engañar incluso a los sistemas de aprendizaje profundo que hayan ganado una de esas competiciones.[9]

Por ejemplo, AlexNet, el sistema que hizo estallar la disciplina en el concurso de 2012, puede llegar a concluir, bajo engaño y con gran confianza, que las imágenes de un autobús escolar, una mantis religiosa, un templo y un shih tzu son avestruces.[10] Los investigadores los llaman «ejemplos antagónicos», y se obtienen alterando estratégicamente unos pocos píxeles en las imágenes —tan pocos que los cambios no son en absoluto perceptibles para el ojo humano—. Para las personas, las imágenes siguen siendo exactamente iguales que los originales.

Los denominados ataques antagónicos tampoco son exclusivos de AlexNet. De hecho, los sistemas de aprendizaje profundo que llegan a mostrar un rendimiento notable en el reconocimiento de imágenes no comprenden lo que perciben. Por tanto, resulta sencillo exponer la fragilidad de su método. Otros experimentos han rebajado de manera drástica su rendimiento limitándose a incluir objetos en segundo plano, algo que los humanos pueden ignorar sin problemas pero que resulta problemático para los sistemas de aprendizaje automático. En otros experimentos, imágenes que parecen extraídas de la estática televisiva —montajes aleatorios de píxeles blancos y negros— han engañado a los sistemas de aprendizaje profundo, que los catalogaron como imágenes de armadillos, guepardos o ciempiés. A medida que la IA contemporánea progresa, esos defectos evidentes dan fe en realidad de la profundidad de conocimiento y contexto que posibilita incluso la percepción visual. Peirce llevaba razón, en otras palabras, acerca de la azalea que vio aquella hermosa mañana de primavera, o acerca de ver cualquier otra cosa:

Realizo una abducción cada vez que expreso cualquier cosa que haya visto en una frase. La verdad es que el entramado al completo de nuestro conocimiento es

un fieltro opaco de hipótesis puras confirmadas y refinadas a través de la inducción. No se puede realizar el menor avance en términos de conocimiento más allá de la fase en que nos quedamos con la mirada perdida sin realizar una abducción a cada nuevo paso.[11]

El origen de la inteligencia, pues, es conjetural o abductivo, y tiene una importancia capital que dispongamos de un armazón conceptual potente desde el que ver los hechos o los datos. Cuando un agente inteligente (persona o máquina) genera una conjetura, explica Peirce, la inferencia descendente, como la deducción y la inducción, deja a las claras las implicaciones de esa conjetura (deducción) y nos proporciona un medio para ponerla a prueba frente a la experiencia (inducción). Las diversas lógicas encajan entre sí: «La deducción demuestra que algo debe ser; la inducción demuestra que algo se encuentra en realidad operativo; la abducción apenas sugiere que algo podría ser».[12] Y, sin embargo, es ese «podría ser» —la abducción— lo que hace que el pensamiento prenda en los entornos del mundo real.

El carácter revisable de la abducción ayuda a explicar el rol capital que desempeña en la comprensión del lenguaje natural, y no solo en los ámbitos de la caza y de la labor detectivesca. Nuestra comprensión de lo que se dice en el lenguaje cotidiano se halla sujeta de manera constante a actualización y revisión. Piensa en este fragmento: «Raymond vio una cachorrilla en la ventana. La quería». El pronombre «la» con toda probabilidad se refiere a la «cachorrilla» (los lingüistas lo consideran un ejemplo de anáfora pronominal o «repetición»). Las dos frases surgen de la nada y no disponemos de más información sobre Raymond, pero el sentido común nos recuerda que la gente, por lo general, desea tener cachorros antes que ventanas, y que a menudo miramos en los escaparates*c objetos que puedan tener algún interés para nosotros. Pero la inferencia de que «la» se refiere a «cachorrilla» no está sujeta a ninguna certeza. Cambiemos el ejemplo añadiéndole contexto y la inferencia se volverá incorrecta: «Raymond rompió la ventana de su casa. Salió a comprar una nueva. Se dijo a sí mismo que al verla sabría cuál era la adecuada. Había una hermosa contraventana de oferta. Raymond vio una cachorrilla en la ventana. La quería. Era la ventana adecuada para él».

Quizá sea un ejemplo forzado, pero no hay nada erróneo en él. Raymond podría pertenecer a ese tipo de personas que dan muchas vueltas a compras en

apariencia prosaicas. Podría incluso ser proclive a tomarse la aparición de una cachorrilla como señal de que ha de comprarse una ventana en concreto, quizá fruto de alguna superstición. La cuestión es que, si vemos el pronombre «la» como una señal, este puede apuntar o referirse a diferentes nombres según los cambios del contexto. Todas las estrategias que han surgido en la IA hasta la fecha han fracasado a la hora de tener en cuenta adecuadamente esos ejemplos.

Los investigadores se refieren a la deducción como «inferencia monótona» porque sus conclusiones son permanentes —cuando el sistema de IA deduce una conclusión, esta se añade de manera automática al almacén de conocimientos del sistema—. No obstante, la comprensión del lenguaje es no monótona (requiere de inferencias revisables). La información novedosa que proporcionen las frases sucesivas puede obligar a realizar cambios en la interpretación inicial: «En ese momento, Raymond supo que aquella era la ventana que necesitaba». Para captar la esencia de una narración tenemos que comprender la manera en que cada nueva frase afecta a la interpretación de las anteriores. Esto se suma a la abducción, que, al fin y al cabo, es conjetural y está sujeta a revisiones desde el primer momento.

En la IA clásica (la IA anterior a la red), los investigadores probaron diferentes cosas a fin de extender la inferencia y hacerla impugnable. El enfoque más común fue ampliar la deducción. El trabajo en el llamado «razonamiento no monótono» alcanzó su pico en los años ochenta y noventa, pero desde entonces ha quedado bastante abandonado, en parte porque las extensiones de la deducción que le prestan la flexibilidad necesaria para que comprenda el lenguaje funcionan solo con ejemplos «de juguete», que no resultan útiles en el mundo real. Un ejemplo clásico se encuentra en el siguiente razonamiento: «Si x es un ave, x puede volar. X es, de hecho, un ave. Así que puede volar. ¡Un momento! Es un pingüino. Los pingüinos no pueden volar. Por consiguiente, x no puede volar (después de todo)». Hay sistemas de razonamiento revisables que permiten razonamientos como este, pero son famosos por mostrarse intratables en casos generales (es decir, que no son revisables informáticamente) y jamás han crecido en escala para tratar con las inferencias complejas que resultan necesarias para interpretar el lenguaje cotidiano que aparece en los artículos de noticias. Los sistemas de inferencia no monótona solo funcionan en escenarios artificiosos generados en el laboratorio.

Sin embargo, por mucho que esos sistemas puedan crecer en escala, el problema central de la deducción sigue siendo las limitaciones que conlleva preservar la

verdad —todo ha de ser cierto—. Podemos modificar o rechazar inferencias luego, pero si comenzamos por usar la deducción estaremos perdiendo el tiempo.

Durante los años ochenta y noventa, muchos de los investigadores que trabajaban en la extensión de la IA para hacerla revisable también desarrollaron métodos de base deductiva para la abducción, sobre todo con la programación lógica abductiva (ALP en inglés). Sin hurgar en sus detalles técnicos, una inferencia de ALP es una consecuencia lógica (una inferencia deductiva que preserva la verdad) entre una teoría lógica T (la base de conocimiento) y la verdad de una condicional E → Q, donde E es una explicación de Q, la observación. En efecto, se trata de una manera elegante de eliminar la naturaleza conjetural de las abducciones. No se obtiene ningún poder inferencial, lo cual explica por qué el trabajo en ALP, como las estrategias de razonamiento no monótono, se abandonó hace tiempo y languidece desde entonces. (El problema de la ALP resulta adecuado para demostrar el problema central y general de la inferencia de sentido común, algo sobre lo que volveremos en una sección próxima.)

Cuando intentamos preservar el concepto de abducción de Peirce como conjetura de una hipótesis verosímil acabamos metiéndonos en un «problema» inferencial. En particular, acabamos obteniendo un error de razonamiento. En los estudios de lógica se conoce como «falacia», y ahora mismo vamos a centrarnos en ella.

FALACIAS E HIPÓTESIS

Peirce representó la abducción como una deducción «rota». La manera más sencilla de entenderlo es retomando la regla del modus ponens para la deducción directa:

A → B (Conocimiento)

A (Observación)

———

B (Conclusión)

Y la tabla de verdad:

A →	B	Conclusión
Verdadero	Verdadero	Verdadero
Verdadero	Falso	Falso
Falso	Verdadero	Verdadero
Falso	Falso	Falso

El problema se encuentra en la segunda fila: A es verdadero pero B es falso, lo que lleva a que la inferencia de A → B sea falsa. Por ejemplo, si A representa «Está lloviendo» y B representa «Las calles están mojadas», la expresión condicional sostiene que siempre es verdad que, cuando llueve, las calles se mojan. Pero si A representa «Está lloviendo» y B representa «Las calles están secas», el condicional material A → B (si A, entonces B) es falso. Es lo que nos dice la segunda fila de esta tabla de verdad.

Fíjate en la tercera fila: si no está lloviendo pero las calles están, de hecho, mojadas, bueno, las calles siguen estando mojadas, así que la regla sigue preservando la verdad. Pero en cualquier momento en que no esté lloviendo y las calles no estén mojadas la regla estará equivocada, así que el resultado de aplicarla será falso. Esto es lógica estándar y se la llama «proposicional» porque las variables representan afirmaciones completas, o proposiciones.

La lógica proposicional se desarrolló hace mucho tiempo y se ha demostrado que no contiene errores. Es completa, lo que quiere decir que sus verdades se pueden probar (siguiendo sus reglas) y que todo lo que se pueda probar es también cierto. También es consistente, porque no puede probar una contradicción. Si A es verdad, «no A» tiene que ser falso, y el sistema no puede demostrar lo segundo si demuestra lo primero. En el mundo simplificado de la lógica proposicional nada se puede torcer. Puedes obtener la verdad de todas las proposiciones que se expresen en él y nunca podrás recibir un sinsentido —pensar, pongamos, que está lloviendo y que no está lloviendo a la vez—. El sistema es consistente.

Ahora piensa en la siguiente falacia, que no se admite en la lógica proposicional ni en la deducción en general:

Afirmación del consecuente

A → B

B

———

A

Los lógicos consideran que esta forma de argumento es un ejemplo de «afirmación del consecuente», porque B es el consecuente de la regla (A se llama «antecedente») y B viene dado como segunda premisa, el caso o hecho observado. Es evidente, no obstante, que al utilizar el consecuente como caso nos estaremos equivocando. Se tratará de un error de razonamiento. Esto se deriva de que A podría ser falso, así que no podemos concluir que sea cierto en todas las situaciones, tal y como requiere la deducción. Puesto que la deducción ha de preservar la verdad, la inferencia es inválida. Estamos ante una falacia.

Resulta más sencillo ver esto asignando a A y B sendas frases (proposiciones):

Cuando llueve, las calles están mojadas.

Las calles están mojadas.

Por consiguiente, está lloviendo.

El argumento no es válido porque, aunque aceptemos la verdad de las premisas, su conclusión no es necesariamente cierta: la calle podría estar mojada por otros motivos (por ejemplo, podría haber estallado una boca de incendio). Afirmar el consecuente es una mala deducción porque se trata de una suposición. La visión de la forma lógica de la abducción como una variante de mala deducción ayuda a explicar por qué ha sido ignorada históricamente en los estudios sobre la razón, y también por qué se ha resistido a los métodos mecánicos como los que encontramos en la IA. ¿Cómo hemos de incorporar una regla incorrecta?

En efecto, la propia formulación que realizó Peirce sobre los tipos de inferencia deja a las claras, precisamente por el motivo que acabamos de comentar, que no se puede transformar la abducción en un tipo de deducción. Él se sirvió de silogismos, expresados en proposiciones de lenguaje natural:

DEDUCCIÓN

Todas las alubias de esta bolsa son blancas.

Esas alubias pertenecen a esta bolsa.

Por consiguiente, estas alubias son blancas.

INDUCCIÓN

Estas alubias pertenecen a esta bolsa.

Estas alubias son blancas.

Por consiguiente, todas las alubias de esta bolsa son blancas.

ABDUCCIÓN

Todas las alubias de esta bolsa son blancas.

Estas alubias son blancas.

Por consiguiente, estas alubias pertenecen a esta bolsa.

Si lo pasamos a la lógica proposicional, tendremos:

Deducción

$A \rightarrow B$

A

______

B

Inducción

A

B

———

$A \rightarrow B$

Abducción

$A \rightarrow B$

B

———

A

En otras palabras, por su propia naturaleza, la abducción no puede convertirse en una forma extendida de deducción, ya que su forma lógica (su esencia) es una falacia deductiva grave, que quebranta esa voluntad de preservar la verdad que presenta por naturaleza la inferencia deductiva. Y tiene sentido, puesto que, tal y como argumentó Peirce, parte de una conjetura o suposición, que por definición puede ser errónea.

De hecho, los tres tipos de inferencia de este armazón se encuentran muy definidos: ninguno de ellos puede transformarse en otro, lo cual implica que, si la inferencia inteligente requiere la abducción, no podremos llegar a ella a través de la deducción o de la inducción. Esta observación tiene una importancia

capital para el trabajo en IA. Si la deducción es inadecuada, y si la inducción es inadecuada, pasamos a necesitar una teoría de la abducción. Puesto que no disponemos de ella (aún), podemos concluir desde ya que no nos encontramos en la senda de la inteligencia artificial general.[13]

He señalado con anterioridad que la conjunción del supuesto de frecuencia y de la restricción empírica eliminan la inducción como estrategia completa de cara a una inteligencia artificial general. La dependencia en las frecuencias de datos nos conduce al pavo de Russell, cuya confianza en que el granjero se interesa por su bienestar no deja de crecer a medida que las observaciones de sus cuidados se van amontonando —el día antes de la comida de Navidad es, inductivamente, el de mayor probabilidad de que sus convicciones sean correctas, ya que es el que cuenta con un mayor apoyo inductivo—. Y la restricción empírica es un límite estricto para el conocimiento o la teoría que podemos ofrecerle al pavo. Si le decimos: «Presta atención a la manera en que te trata el granjero», el pavo será cada vez más feliz hasta el día de su muerte. Pero, puesto que la observación «Me han puesto sobre la tabla de cortar» no aparece entre los datos en ningún momento previo a (t = Navidad), no es posible suministrársela al modelo o teoría del pavo por inducción pura.

En el aprendizaje automático, esto significa que el único conocimiento que podemos suministrarle a un sistema es el que se puede recuperar entre los datos de manera puramente sintáctica. Esto se ha visto como una virtud, como en el caso del sistema Atari de DeepMind, pero implica el mismo tipo de ángulo muerto que atenaza al pavo de Russell —que no conoce aquello que no puede observar en los datos—. Esto da pie a predicciones fallidas, como los defectos del reconocimiento de imágenes, y también es responsable de la peculiar fragilidad de los sistemas modernos, donde las alteraciones en apariencia triviales en los píxeles degradan su rendimiento en juegos y otras tareas.

Podemos apañar algún sistema inductivo, basado exclusivamente en datos, incluyéndole más datos... hasta cierto punto. Pero las excepciones, las observaciones atípicas y todo tipo de sorpresas forman parte integral del mundo real. La estrategia de exponer los sistemas de aprendizaje supervisado a excepciones probables, tal y como se hace en el trabajo en curso con los coches sin conductor, es una tarea digna de Sísifo, porque, por su propia naturaleza, las excepciones no se pueden predecir por completo. Hace falta un enfoque nuevo, esencialmente abductivo. Mientras tanto, estamos atrapados en las regularidades observables, las «testarudas» regularidades que pueden automatizarse.

Judea Pearl, que tocó bien este punto con la metáfora de la «escalera de causalidad», dice del aprendizaje automático y de la estadíst ica que representan un ejercicio de «encaje de datos en una curva» (y técnicamente es así), lo cual coincide con su primer peldaño, el de la asociación. En él podemos hacer preguntas sobre la correlación como «¿Qué me dice un estudio sobre los resultados de las elecciones?». Y podemos usar correlaciones entre los movimientos de una partida y las victorias resultantes para diseñar sistemas modernos de juegos como AlphaGo. Pero no podemos extraer informaciones causales sobre el mundo de las asociaciones que surjan entre los datos, así que resulta imposible formular, y mucho menos responder, a explicaciones que se pregunten el porqué o el cómo. El conocimiento causal forma parte de nuestra comprensión del mundo, basada en el sentido común, y explica, por ejemplo, que podamos ver los datos como efectos o pistas de causas anteriores que contribuyan a nuestra comprensión.

Es importante que en el peldaño más alto de la escalera de Pearl se encuentren los contrafactuales, donde realizamos preguntas hipotéticas cuyas respuestas no existen en ningún conjunto de datos (por definición, porque preguntamos por cosas que no han sucedido). Nos imaginamos mundos. En esos contrafactuales, parte integrante de la inteligencia humana, los datos son inútiles para ayudarnos a determinar, pongamos, si Kennedy seguiría vivo en el caso de que Oswald no hubiera nacido nunca, o si el pavo estaría a salvo si al granjero le hubieran regalado otro pavo, o si las calles continuarían estando mojadas si un autobús no hubiera golpeado la boca de incendio al pegar un volantazo. La imaginación incluye inferencias que no existen dentro de un conjunto de datos. Y el acto de imaginar requiere, más que ninguna otra cosa, de la conjetura. La abducción es una inferencia que yace en el centro de toda la inteligencia.

Aunque en estos últimos tiempos los investigadores parecen haberlo olvidado, a lo largo de la mayor parte de la accidentada historia de la IA, su desafío central ha consistido en adquirir y utilizar un conocimiento de sentido común acerca del mundo. El sentido común obliga a tener una comprensión muy rica del mundo real, lo cual de manera amplia consta de dos partes: primero, los sistemas de IA deben de algún modo adquirir conocimientos cotidianos (en grandes cantidades); y, segundo, deben poseer algún tipo de capacidad inferencial para hacer uso de él.

Hector Levesque, científico de IA de la universidad de Toronto, plantea un buen argumento en su libro de 2018 Common Sense, the Turing Test, and the Quest

for the Real AI [«Sentido común, el test de Turing y la búsqueda de la IA real»]: en ausencia de una teoría sofisticada del conocimiento, nuestros esfuerzos por conseguir una inteligencia artificial general (una IA «real») están condenados al fracaso.[14] Su análisis sobre la necesidad de un conocimiento de sentido común da en el clavo, pero al parecer cae en la ya tradicional trampa de entender la inferencia como una extensión (todavía desconocida) del razonamiento deductivo.

En gran medida, Levesque intenta resucitar un campo de la IA que en su día gozó de prominencia, que se conoce como «representación del conocimiento y el razonamiento» (KR&R en inglés) y que abordó de frente los problemas de conocimiento e inferencia en los sistemas inteligentes. La parte de razonamiento de la KR&R incluye consideraciones sobre la inferencia, y los problemas que acabamos de repasar ya han sido más o menos descubiertos en su terreno, pero, por desgracia, se han quedado sin solucionar. En un artículo de 2013 titulado «On Our Best Behavior» [«Para portarnos como nunca»], Levesque señalaba que las ampliaciones (y los parches) de la deducción flaquean porque son «problemas intratables», término de las ciencias de la informática que significa que las soluciones no se pueden computar en tiempo real (si es que llegan a ser computables):

Incluso el conocimiento más básico e infantil parece recurrir a una amplia gama de constructos lógicos. Causa y efecto y falta de efecto, contrafactuales, cuantificadores generalizados, incertidumbre, las convicciones de otros agentes, sus deseos e intenciones, etc. Y, sin embargo, el razonamiento simbólico que se eleva por encima de esos constructos parece resultar demasiado exigente en términos informáticos.[15]

Los problemas intratables representan una de las pistas de que el enfoque mismo está equivocado. Una pista de mayor peso sería que la deducción no tenga encaje en la lógica de la abducción. Ese carácter irreductible implica que el problema es de fundamentos, más allá de los problemas de coste de la computación.

Ambos aspectos de la KR&R —es decir, tanto la representación como el razonamiento— son necesarios en la actualidad. La representación del

conocimiento de sentido común en la máquina se ha revelado especialmente compleja, por decirlo con suavidad. Pese a las décadas de trabajo, nadie se ha acercado siquiera a introducir en una máquina el conocimiento básico suficiente como para conferirle el poder de actuar en escenarios del mundo real como serían entender el lenguaje común y corriente, o desplazarse alrededor de una casa o por una calle bulliciosa.[16] Es evidente que el conocimiento y el razonamiento están ligados, porque no podemos inferir lo que ignoramos, y no podemos hacer uso del conocimiento que tenemos sin una capacidad de inferencia adecuada. Yo me refiero a esos problemas como «el pozo sin fondo del conocimiento» y «el motor de la inferencia mágica».

Irónicamente, el propio viejo esfuerzo de la programación lógica abductiva demuestra con acierto sus escollos clave. Supón (de nuevo) que T es una base de conocimiento del sistema S de IA, que realiza la (sorprendente) observación Q, según la cual una explicación (o causa) E se encuentra en T, y T implica que E → Q, donde implicar es una inferencia que preserva la verdad con fuerza (técnicamente, la consecuencia lógica es más potente que la implicación material «→», ya que cada proposición de T debe hacer también que E → Q, y viceversa). Así, T es la «teoría del todo» de S. Primera pregunta: ¿cómo logramos introducir todo el conocimiento necesario en T? Segunda pregunta: puesto que E → Q es solo la condición material que hemos visto arriba, ¿cómo es posible que la verdad de E → Q constituya una inferencia abductiva de la conjetura plausible (comprobable) de, en este caso, E para la observación de Q? En otras palabras: ¿cómo llegamos a una teoría T y cómo la usamos de manera abductiva? Lo único que se necesita para explicar estos problemas es que el pavimento esté mojado.

El departamento estadounidense de Defensa, a través de su brazo de investigación de alta tecnología, la Agencia de Proyectos de Investigación Avanzados de Defensa (DARPA en inglés), invirtió en una ocasión una gran suma de dinero para la elaboración de unas bases amplias de conocimiento de sentido común. Expertos formados en lógica e informática (a decir verdad, yo fui uno de esos expertos) fuimos alimentando aquellos sistemas informáticos a cucharadas, con proposiciones comunes y corrientes del tipo «Los seres humanos vivos tienen cabeza», y «Los aspersores disparan agua», y «El agua moja», etc. La idea era que los sistemas de IA, armados con montones de aquellos datos de sentido común, podrían beber de ellos para inferir cosas acerca del mundo. Se esperaba que esos sistemas no necesitaran de la inferencia abductiva, porque los sistemas repletos de conocimiento podrían utilizar

enfoques más sencillos y comprensibles para llegar a la inferencia. Puesto que «Cuando conectan los aspersores, la zona circundante se moja» es una proposición tan cierta como otras similares acerca de la lluvia, los motores de inferencia de los sistemas podrían, al fin y al cabo, quedarse en la deducción, ya conocida. Los sensores (o la entrada de texto) informarían a los sistemas de IA de la llegada de un tsunami y la deducción inferiría que las calles iban a mojarse.

Pero ante aquel enfoque afloraron varios desafíos de importancia. El primer problema era evidente, o debería haberlo sido: la mayor parte de lo que sabemos es implícito. Llevamos ese conocimiento a la consciencia, volviéndolo explícito, solo cuando lo requieren las circunstancias, como cuando nos sorprendemos o tenemos que reflexionar sobre algo de manera deliberada.

Eso nos conduce al segundo problema, «la punta del iceberg». Nuestro conocimiento implícito puede resultar necesario para alguna inferencia u otra, pero su suma total es amplísima. El conocimiento base de una persona normal y corriente es increíblemente grande, e introducirlo y representarlo en un ordenador se convierte en una labor pantagruélica.

Lo de alimentar a un ordenador con cucharadas de sentido común resultó ser un proyecto filosófico para toda la vida; sacábamos a la luz conocimiento de sentido común del tipo «Si viertes líquido en un contenedor de cristal sin agujeros y con una sola abertura, este se llenará». O que «Los seres humanos vivos tenemos cabeza», o que «La carretera es una vía de superficie dura que se dedica al tráfico de vehículos». Los investigadores habían asumido que los ordenadores acabarían «pillándolo», pero lo que pasó fue que el proyecto se volvió interminable.

Imagina un sistema de IA construido solo para responder de manera inteligente al tema rematadamente simple y aburrido de la humedad. Él mismo necesitaría una base de conocimientos inmensa. Cualquier máquina que, pongamos, pudiera mantener una conversación realista solo acerca de ese tema necesitaría conceptos como el de los aviones de extinción de incendios que transportan agua (y no gasolina, aunque eso también serviría para mojar las calles), el de las pistolas de agua, los juegos infantiles, etc. Piensa que le hicieran al sistema una pregunta sencilla, similar a las del test de Turing: «Era un día caluroso, no había una sola nube en el cielo. Llamaron a los bomberos para que fueran a cerrar una boca de incendio estropeada. La calle principal estaba mojada, y las alcantarillas se habían obstruido con los restos que había arrastrado el agua. ¿Por qué estaban

mojadas las calles?». Si la base de conocimiento del sistema de IA no dispone de bomberos y bocas de incendios, no cabe esperar que responda a la pregunta. Y considera que añadamos unas frases más a la pregunta: «Pero eso no fue lo que mojó las calles. El chorro de la boca de incendio se proyectaba sobre el escaparate del colmado. Una tormenta tremenda cayó sobre la zona justo antes de que los bomberos hicieran acto de presencia. ¡Todo estaba mojado!». ¿De qué sirven todos esos conceptos de la base de conocimiento? ¿Y la deducción?

DEL CONOCIMIENTO COMPUTACIONAL COMO POZO SIN FONDO

En eso radica el problema del «pozo sin fondo»: llenar una base informática de conocimiento con afirmaciones expresadas en forma de proposiciones (dentro de una lógica) es una tarea interminable. No podemos resolver siquiera problemas simples de sentido común, como razonar sobre lo que sucede en una calle o barrio de una ciudad, sin codificar de manera eficaz inmensos volúmenes de conocimiento en apariencia irrelevante.

También está el problema de la representación. Las bases de conocimiento, igual que las bases de datos relacionales, deben organizarse y estructurarse, de modo que los fragmentos de conocimiento necesarios para hablar sobre calles mojadas, o sobre la ausencia de Kate en la cafetería, o sobre las latas de atún a medio comer, o sobre lo que sea, se encuentren disponibles para realizar inferencias. No nos servirá que un sistema de IA se ponga a calcular la posición de Marte en el cielo nocturno cuando le hemos preguntado si alguien se comió el atún con cuchara. Ese es el problema de la relevancia, que se debe frustrar de manera astuta, con estrategias para la representación del conocimiento que lo vuelvan accesible y disponible. Por lo general, los investigadores intentan «resolver de antemano» los problemas de relevancia agrupando conocimientos relacionados entre sí, de modo que las reglas (como modus ponens y muchas otras) cubran lo que sea necesario.

En general, la base de conocimiento se organiza de manera jerárquica utilizando vínculos de tipo «es-un», indicando que algo es una instancia de otra cosa, como si dijéramos que un portátil es un ordenador personal, que es un tipo de ordenador, que es una forma de tecnología digital, etc. En las bases de

conocimiento amplias, las jerarquías conforman la columna vertebral de grupos enciclopédicos de temas sobre las cosas que con frecuencia encontramos en el mundo. Junto con todos esos vínculos «es-un», se pueden introducir otros predicados (esto es, vínculos entre conceptos) que expresen otros parentescos importantes, como las relaciones parte-todo (meronimias) que existen entre los conceptos.[17] Los lenguajes de representación del conocimiento han evolucionado desde los primeros años de la IA para facilitar el desarrollo de las bases de conocimiento, usando vocabularios simplificados y hechos a medida para expresar relaciones jerárquicas.

Durante la década de 1990, por ejemplo, en la red hubo grupos de trabajo que desarrollaron marcos de descripción de recursos (RDF en inglés) escribiendo «tripletes» (frases con sujeto-predicado-objeto directo) cuyo predicado podía ser «es-un», «parte-de» o cualquier otra cosa que se considerara de utilidad para construir la base de conocimiento. Los RDF ayudaron a que las bases de conocimiento se convirtieran, en esencia, en enciclopedias de computación (según las palabras del fallecido investigador de IA John Haugeland) donde las bases de conocimiento de proyectos más amplios llegaron a disponer de miles de tripletes. Los investigadores de IA esperaban que la facilidad de uso animara a que incluso los neófitos realizaran tripletes —sueño que articuló Tim Berners-Lee, el creador del HTML—. Berners-Lee la bautizó como la «web semántica» porque, con aquellas páginas web convertidas en proposiciones RDF legibles por las máquinas, los ordenadores sabrían qué significaba todo. Los ordenadores iban a leer de manera inteligente toda la red. Los investigadores de IA vendieron las bases de conocimiento como el fin de los sistemas frágiles que utilizaban solo estadísticas —porque, al fin y al cabo, las estadísticas no bastan para llegar a la comprensión—. La web semántica y otros proyectos de IA centrados en las bases de conocimiento podrían al fin «conocer» el mundo y hacer algo más que resolver sudokus. Se malgastó una enorme cantidad de dinero y esfuerzos en aquel sueño, que nunca llegó a funcionar. El problema del pozo sin fondo sigue acompañándonos, porque la tarea es interminable.

Pero, puesto que por lo general el conocimiento se organiza por contenidos, de modo que el sistema de IA pueda acceder con mayor rapidez a él (ya que, después de todo, este no se puede pasar el día entero reflexionando sobre por qué están mojadas las calles cuando le plantean una pregunta), los ingenieros se descubrieron añadiendo cada vez más información «extraña» al área de cada tema concreto para abarcar todo lo que pudiera ir surgiendo. Volviendo a nuestro ejemplo, bomberos, bocas de incendio, pistolas de agua, aviones de extinción,

lluvia, nieve, inundaciones y demás..., todo ello tenía que estar conectado de algún modo para que el sistema tuviera la esperanza de llegar a contestar las preguntas más básicas acerca de unas calles mojadas. Pero esa estrategia no tarda en dejar de tener sentido; frustra la intención original de organizar el conocimiento de manera natural, por temática. Mientras que las bocas de incendio podrían formar parte de la descripción de, pongamos, una típica manzana urbana con sus calles, lo más probable es que no sea así con unos aviones enormes contra incendios o la descripción de un juguete infantil que dispara agua. En otras palabras, el esfuerzo mismo por organizar el conocimiento de modo que sea accesible informáticamente de cara a un razonamiento en tiempo real deja fuera de manera inevitable objetos que serán necesarios en un escenario u otro. No parece que exista ninguna receta para construir una base de conocimiento para la inteligencia —al menos, ninguna que nosotros conozcamos, concebida de antemano, para llenarla de lenguajes lógicos como los RDF o cualquier otra cosa.

La prueba definitiva del carácter irresoluble del problema del pozo sin fondo se fue manifestando de forma gradual, cuando los proyectos ya estaban irremediablemente en marcha. Los investigadores se habían dado a proyectos masivos de desarrollo del conocimiento base, y se habían destinado miles de horas y personas a rellenar enciclopedias que se pudieran aprovechar para dominar el problema de sentido común que tenía la IA. No obstante, cuanto más «conocimiento» se les añadía, más posibilidades tenían los sistemas de equivocarse por completo. Un sistema que solo sepa cosas sobre la lluvia y la humedad quizá responda bien a una pregunta, pero, al introducirle docenas de escenarios diferentes, en general, el sistema heredaba cada vez más maneras de errar. «Llamaron a los bomberos porque la tormenta había inundado varios parques» se refiere en realidad a la lluvia, que ha provocado la inundación. Pero la referencia a los bomberos podría despistar a un sistema que «supiera» de bocas de incendio que se rompen y que pueden llegar a inundar una calle. Así, a diferencia del pensamiento humano, donde en general cuanto más sabemos más poder tenemos para realizar inferencias útiles, los proyectos amplios de base de conocimientos corren siempre el peligro de generar inferencias sin sentido con todo su «conocimiento» adicional. Es evidente que no basta con las proposiciones, los hechos y las reglas que conforman el sentido común; la manera en que funciona la misma inferencia de sentido común también tiene su importancia.

En retrospectiva, los primeros intentos de otorgar «sentido común» a los

sistemas de IA fueron dos proyectos que se disfrazaron como uno solo. El conocimiento es un requisito evidente, pero la inferencia también lo es. Conocemos solo una parte de ella; la manera en que usamos lo que sabemos para actualizar nuestras convicciones es otra bastante diferente. Pero las estrategias para llegar a la inferencia informática son limitadas. Tal y como señaló Peirce, tenemos tres, y hasta la fecha solo dos de ellas se han reducido a una forma de computación. Cuando afloraron los problemas de relevancia respecto a la deducción pura, los investigadores probaron con varios planes ad hoc.

En los años setenta por ejemplo, Roger Schank desarrolló en Yale una aproximación al sentido común basada en «guiones».[18] Schank argumentó que los pensamientos y acciones cotidianos siguen un guion implícito: tienen un hilo argumental. Un ejemplo paradigmático sería cuando pedimos la comida en un restaurante. Entramos en él, nos sientan (o escogemos sitio nosotros), miramos el menú y, cuando llega el camarero, tras un saludo inicial, pedimos. Todo esto ocurre de manera secuencial y hasta cierto punto se puede anticipar y planear. Schank asumió que algunas de las actividades del mundo real podían guionizarse de esa manera y desarrolló algunos sistemas que utilizaban guiones programados para interactuar con la gente en tareas como la de pedir la comida. Sin embargo, en el mundo real no hacemos más que salirnos del guion. Necesitamos el sentido común para pedir la comida, o para hacer una reserva cuando sucede algo inesperado, como que tarden demasiado en venir a atendernos. Es posible que haya un letrero en la puerta de entrada para indicar a los clientes que deben usar la puerta lateral.

El trabajo de Schank con los guiones dejó de tener la consideración de enfoque serio para resolver el problema de la IA con el conocimiento de sentido común, pero sigue resultando instructivo porque muestra con claridad la separación que existe entre conocimiento e inferencia. Podríamos suponer que se le puede transmitir a un ordenador todo el conocimiento relevante a la hora de pedir la comida —que todos los conceptos y predicados que describen el restaurante y el momento de pedir la comida se encuentran en la Base de Conocimiento de Restaurantes—. Pero las alteraciones, los sucesos inesperados o una simple conversación con el camarero no tardarán en confundir al sistema basado en guiones. De manera invariable, este necesitará saber mucho más acerca del mundo de lo que aparece en la Base de Conocimiento de Restaurantes.

Resultaría tentador intentar suministrarle todas las cuestiones extrarrestaurantiles a través de un «metaconocimiento» basado en todo tipo de cosas de sentido

común: el tiempo, los partidos que pasan en la televisión, etc. Pero, puesto que no habrá enlaces directos a todo ello (a causa del problema que he mencionado antes: que no podemos resolver de antemano todo el conocimiento que puede volverse importante), el sistema debe de algún modo saltarse el guion de conocimiento y los planes relacionados con el tema. Pero, para hacerlo, necesita tener algún mecanismo de inferencia que sepa lo que está pasando —dónde debe mirar y en busca de qué—. Y, para ello, antes debemos resolver un problema diferente del conocimiento per se. Necesitamos un mecanismo de inferencia que no sea deductivo (ni inductivo) y que sí se muestre flexible y esté dotado de sentido común. Lo cual nos lleva de nuevo a la inferencia.

LOS MOTORES DE INFERENCIA MÁGICA: EL PROBLEMA DE SELECCIÓN PARA LA IA

Si tenemos una norma, pongamos que A → B, y observamos un hecho B (sorprendente), podemos lanzar la hipótesis de que A es cierto, porque explicaría B (ya que se sabe que A → B). Por tanto, si «Cuando está lloviendo, las calles están mojadas» es una proposición conocida y verdadera, y vemos que las calles están mojadas, sin saber nada más podríamos suponer que ha estado lloviendo. En efecto, dentro de un contexto inesperado, será lo más probable.

Pero un sistema que también observe C, digamos «El cielo está despejado», debería ignorar la regla A → B. ¿Por qué? Pues porque ya no es relevante. De hecho, si recordamos que A es una variable, puede significar un montón de proposiciones diferentes: «Está lloviendo» o «Los niños están jugando con pistolas de agua», etc. De hecho, podemos extender A para que incluya un conjunto de proposiciones como estas: A = {Lluvia, Boca de incendio, Aspersores, Pistolas de agua, Tsunami...}. Entonces, la tarea consistirá en seleccionar, entre los miembros del conjunto A, bajo el antecedente de la regla A → B, la proposición verdadera que resulte más relevante para la observación B. Entender la relevancia es inevitable; conocer B no nos dice nada acerca de qué miembro del conjunto A debería usarse en A → B. Debemos elegir al que consideremos el miembro más relevante del grupo.

De ese modo, la inferencia deductiva entre la observación A y la conclusión B es

cierta pero demasiado sencilla. Queremos observar el efecto e inferir la causa: las calles mojadas y su porqué, o cómo. Todo eso forma parte de la inteligencia normal, de lo que Marcus y Davis definen como disponer de una imagen elocuente del mundo, en la que se vuelve crucial tener el conocimiento de qué causa qué. Adquirir y utilizar ese conocimiento resulta complicado, porque la mayoría de los hechos del mundo real admiten numerosas causas posibles. El problema de selección consiste en encontrar la causa operativa o la mejor o la más verosímil, dadas todas las posibilidades, reales o imaginarias. Por lo tanto, el problema central de la inferencia abductiva automatizada puede reformularse como este problema de selección, que ayuda a exponer la dificultad inherente a la inferencia requerida, pero en última instancia se trata del mismo problema. Para abducir tenemos que resolver el problema de selección que existe entre las causas o factores enfrentados, y para resolver ese problema tenemos que captar de alguna manera lo que resulta relevante en una situación u otra. El problema es que nadie tiene ni idea de cómo hacerlo. Nuestras inferencias actuales son a menudo suposiciones que consideramos relevantes o verosímiles, pero no son deducciones ni inducciones. Ese es el motivo por el que, desde el punto de vista de la IA, parecen mágicas.

Peirce ofreció lo que él denominó una «explicación vaga» sobre la irrazonable precisión de nuestras conjeturas: «Pienso que no puede existir ninguna duda razonable sobre el hecho de que la mente del hombre, al haberse desarrollado bajo la influencia de las leyes de la naturaleza, piensa con naturalidad siguiendo los patrones de esa misma naturaleza».[19] Estimó, quizá de manera conservadora, que «el billón (a saber, un millón de millones)*d de hipótesis que podría conjeturar una criatura fantástica explicarían cualquier fenómeno dado».[20] También tenía en la cabeza una base de conocimiento vinculado de algún tipo: «Ya que estos fenómenos sin duda se hallarían más o menos interconectados en la mente de esa criatura con otro millón de fenómenos diferentes (puesto que no se limitaría a los acontecimientos contemporáneos)». Concluyó, con una suerte de rechazo, que «no llevar[ía] esa idea más allá», porque una criatura tal, en ausencia de un «patrón natural», se enfrentaría a la posibilidad irrealizable de tener que conjeturar correctamente por azar.[21] Lamentablemente, la IA no puede desdeñar el problema; se trata precisamente del problema que debe solucionar.

PENSAR RÁPIDO, PENSAR DESPACIO

La idea de que unos instintos innatos impulsan parte de nuestro pensamiento tiene un largo historial y se presenta bajo una apariencia moderna en obras, por ejemplo, del premio Nobel Daniel Kahneman. En su best seller de 2011 Pensar rápido, pensar despacio, Kahneman lanzó la hipótesis de que nuestras mentes pensantes constan de dos sistemas primarios, que denominó Tipo 1 y Tipo 2. El pensamiento de Tipo 1 es rápido y reflejo, mientras que el pensamiento de Tipo 2 consume mayor cantidad de tiempo e implica cálculos deliberados.[22] La percepción de una amenaza, como la de un hombre que se te acerca con un cuchillo en una calle oscura, es un ejemplo del pensamiento de Tipo 1. El pensamiento reflejo, instintivo, toma las riendas en ese tipo de situaciones porque (se supone que) las habilidades del de Tipo 2, de razonamiento meticuloso y deliberado, serán demasiado lentas para que nos salvemos. No podemos ponernos a solucionar problemas matemáticos; necesitamos un juicio inmediato para seguir con vida. El pensamiento de Tipo 2 incluye tareas como sumar números o decidir qué vino irá bien para la cena cuando tienes invitados. En caso de una amenaza potencial, ese pensamiento de Tipo 2 no estará disponible con rapidez y no nos será de ayuda.

En Pensar rápido, pensar despacio, Kahneman argumentó que muchos de nuestros errores al pensar se deben a que permitimos que las inferencias de Tipo 1 se contagien a situaciones en las que deberíamos mostrarnos más concienciados, cautos e inquisitivos. El Tipo 1 sabe cómo desplazar al Tipo 2, lo que a menudo nos conduce a caer en la falacia y el sesgo.

Todo esto es verdad y está muy bien, hasta cierto punto. Pero la distinción entre sistemas de Tipo 1 y Tipo 2 perpetúa un error que han cometido muchos investigadores de IA: pensar que el pensamiento consciente e inteligente es un cálculo deliberado.

En realidad, tanto las cuestiones sobre la relevancia como el problema de selección y el aparato entero de la inferencia basada en conocimiento se encuentran implícitos en los pensamientos de Tipo 1 y de Tipo 2. La distinción de Kahneman es artificial. Si veo que un hombre se me acerca en una calle oscura de Chicago podría inferir con rapidez que representa una amenaza. Pero esa inferencia (presumiblemente, una inquietud de Tipo 2) surge con tanta rapidez que, en el lenguaje, por lo general decimos que hemos percibido una amenaza, o que hemos realizado un juicio instantáneo sobre esa amenaza. La

vimos venir, decimos. Y, en efecto, esta desencadenará una respuesta de lucha o huida, tal y como señaló Kahneman. Pero no es literalmente cierto que hayamos percibido la amenaza sin pensar. Las amenazas que percibimos son inferencias rápidas, desde luego, pero siguen siendo inferencias. No son solo reflejos. (Recordemos la azalea de Peirce.)

La abducción vuelve a representar un papel fundamental en el pensamiento rápido. Pongamos que es Halloween y que entendemos que el hombre que se nos está acercando lleva puesto un disfraz y esgrime un cuchillo falso. O que se trata de Frank, el electricista, que viene por la calle con sus herramientas (que incluyen un cuchillo) y que las sombras se deben a que ha habido un apagón. Todas estas son abducciones, pero suceden con tanta rapidez que no reparamos en que el conocimiento previo ha entrado en juego. El molde de nuestras expectativas hará que lo consideremos como una amenaza o algo inofensivo, incluso cuando pensemos con rapidez. En otras palabras, estaremos conjeturando explicaciones, y eso es también lo que conduce al Tipo 2 de pensamiento. Nuestros cerebros —esto es, nuestras mentes— son generadores de inferencias.

Por decirlo de otro modo, todas las inferencias (veloces o lentas) son noéticas, o basadas en el conocimiento. Nuestra capacidad inferencial está enredada de algún modo en hechos relevantes y fragmentos de conocimiento. La pregunta es: ¿cómo se programa todo eso en una máquina? Tal y como señala Levesque, parece necesario que exista una disciplina, como el razonamiento y la representación del conocimiento de la IA clásica, a fin de progresar hacia la inteligencia artificial general. En la actualidad solo sabemos esto: necesitamos dar con la manera de realizar inferencias abductivas, lo cual requiere de vastos depósitos de conocimiento de sentido común. Aún ignoramos cómo imbuir ese conocimiento en las máquinas y, aunque acabemos averiguándolo algún día, no sabremos implementar el motor de la inferencia abductiva para que utilice todo ese conocimiento en tiempo real y en el mundo real —es decir, no mientras no se dé una innovación conceptual importantísima en la IA.

DEL PROBLEMA DE SELECCIÓN COMO ABDUCCIÓN HIPOCODIFICADA (Y MÁS)

En una colección poco conocida de textos sobre la abducción y la inferencia, el fallecido novelista, semiótico y filósofo Umberto Eco clasificó los tipos de inferencia abductiva según su novedad inherente (y, por tanto, su dificultad informática).[23] Se trata de una clasificación instructiva para la ansiedad que hoy en día despierta en nosotros la IA. Las hipótesis o abducciones sobrecodificadas son, de manera paradigmática, casos de traducción en los que, por ejemplo, el «hombre» español significa «macho humano adulto». Eco señala que incluso estas inferencias en apariencia triviales son automáticas solo en parte, porque el conocimiento previo y el contexto pueden alterar nuestras convicciones. En una cultura extranjera políglota, la palabra «hombre» podría no poner en juego su significado según el diccionario español. Las abducciones sobrecodificadas no han dejado de camuflar las contribuciones de la convicción o hipótesis. Las convicciones de fondo informan a la expectativa del significado de los fonemas (los sonidos de los que están hechas las palabras).

Las abducciones sobrecodificadas requieren de una selección de normas y hechos relevantes que ya se conocen, pero que solo se pueden aplicar a las inferencias dentro de un contexto. Comprender el lenguaje natural es un caso paradigmático de inferencia abductiva subcodificada. «Margaret vio un cuervo en un árbol. Los detestaba» suele interpretarse desde el significado de que Margaret detesta los cuervos, resolviendo que el pronombre «los» se refiere a su antecedente preferido, el común de los cuervos. Pero «Margaret vio un pájaro cantor posarse en su árbol favorito. Pasó a quererlo aún más» traslada la interpretación preferida hacia el árbol —y añadir el aún más reclama un contexto mayor acerca de Margaret, el pájaro cantor y el árbol—. En otras palabras, está subcodificado y hace falta una valoración del sentido de la narración en la que aparecen las dos frases. Los sistemas como Siri o Alexa deben realizar —como mínimo— inferencias abductivas subcodificadas para generar respuestas más contextuales y cargadas de significado. Los ejemplos de dos frases que he ofrecido no son más que una gota en el océano de los problemas para conseguir que esos sistemas funcionen sirviéndose de los métodos existentes.

La abducción no hace más que complicarse. No tarda en alejarse de todas las concepciones conocidas de las inferencias y los cómputos automáticos. Pensemos en las innovaciones o descubrimientos científicos. Los seres humanos inventan lenguajes, conceptos y leyes que explican el mundo. Es la abducción creativa. Las abducciones creativas «dan un salto» por sí solas hacia marcos conceptuales novedosos. Me viene a la cabeza sir Isaac Newton, quien no solo extendió las matemáticas para que describieran tasas instantáneas de variación

en curvas (o la aceleración), sino que otorgó nuevos significados a algunas palabras inglesas a fin de explicar la física. «Gravity» [«gravedad»] se refería antes a algo intenso y serio —como en «gravitas» [«dignidad»]—, y la fuerza de atracción a la que ahora denominamos «gravedad» se entendía como una tendencia o un propósito. Cuando Newton lanzó la hipótesis de que los objetos como las manzanas caían a la Tierra regidos por la gravedad, una fuerza invisible pero que podía describirse matemáticamente, también se dio cuenta de que se trataba, o podía tratarse, de la misma fuerza que tiraba de la Luna a lo largo de su órbita —para que, en efecto, cayera alrededor de la Tierra— y que hacía que el océano se elevara con la marea alta. En esencia, esos conceptos se inventaron en aquel momento.

Turing utilizó la máquina de escribir como modelo para explicar las máquinas universales, o computadoras. Las máquinas de escribir eran conocidas por todo el mundo, pero, de manera improbable, se convirtieron en uno de los más fenomenales inventos científicos y tecnológicos de la historia de la humanidad. Automatizar abducciones creativas como aquella sigue siendo una cuestión incierta, por decirlo con suavidad. Sin embargo, los seres humanos las han hecho y siguen haciéndolas. El descubrimiento de la estructura del ADN por parte de Watson y Crick es otro ejemplo famoso, igual que la teoría de la relatividad de Einstein. Y son innumerables las abducciones creativas menos conocidas que han conducido la ciencia y la innovación por el camino del progreso desde los albores de la sociedad humana.

Las abducciones creativas resultarían menos problemáticas de manera inmediata para la IA si se limitaran a ser los destellos de genio de un Newton, un Turing y demás. Pero los momentos interesantes e importantes de nuestras vidas privadas también suelen ser abducciones creativas. Cuando reconceptualizamos el mundo, por ejemplo, y encontramos un nuevo sentido en los hechos cotidianos, o nos damos cuenta de que las relaciones personales son más importantes que el dinero, o experimentamos una conversión religiosa (o las tradiciones y la fe pierden el poder que tenían sobre nosotros), pasamos a ver las cosas a través de una lente completamente nueva. No se trata solo de que escojamos algo entre una red de posibilidades contextuales, como al resolver un puzle, sino que el sentido es mucho más profundo. En su lugar, pasamos a ver el mundo y sus sucesos y acontecimientos de una manera por completo diferente.

Ese salto creativo tiene lugar a diario y le pasa a mucha gente. Los saltos pueden ser amplios y serios (como con los temas de fe y dudas religiosas) o pequeños y

mundanos. A menudo resultan divertidos. Nos involucramos en una inferencia interesante cada vez que leemos el periódico, o mantenemos una conversación, o transitamos una calle concurrida para ir a hacer la compra. Y las abducciones creativas se asientan detrás de nuestra fascinación y disfrute de la música, el arte, las películas y los relatos. Los detectives —desde los tiempos de Dupin y Sherlock Holmes— nos entretienen porque reconocemos en sus reflexiones el potencial creativo de la mente humana. Nos maravilla la manera en que pueden extender la inferencia inteligente hacia la hazaña del razonamiento.

En resumen, las inferencias abductivas, misteriosas y maravillosas, impregnan nuestra cultura; son, en un sentido amplio, lo que nos convierte en humanos. Es posible que el sueño de una IA capture algún día esos saltos de manera automática, pero mientras tanto deberíamos admitir que la verdad es que no sabemos cómo mecanizar nuestra experiencia.

Todos los caminos conducen a este punto: a la IA le falta una teoría fundamental —la teoría de la inferencia abductiva—. El problema de la comprensión del lenguaje expone ese problema de manera inequívoca, y será un motivo de preocupación central del próximo capítulo. Concluyamos este con algunas observaciones a modo de resumen y recordatorio.

Primero, la inferencia deductiva nos proporciona un cierto conocimiento. Cuando las premisas de un argumento deductivo son verdaderas, y cuando la regla que se ha utilizado para inferir la conclusión es válida (se sabe de ella que contribuye a preservar la verdad), la deducción garantiza que nos desplacemos entre una inferencia real y la siguiente. El problema radica en que las verdades atemporales captan una parte muy pequeña del mundo cotidiano e, incluso cuando tenemos alguna certeza, la inferencia deductiva ignora las cuestiones relacionadas con la relevancia. Así pues, una IA compuesta de deducciones acabaría resultando bastante estúpida y ridícula: por ejemplo, podría concluir que un hombre no se quedará embarazado tras tomarse las píldoras anticonceptivas de su esposa.

Segundo, la inferencia inductiva nos proporciona un conocimiento provisional, porque el futuro podría no parecerse al pasado. (A menudo es así.) Los expertos en lógica dicen de la inducción que es sintética, porque añade conocimiento, sin embargo es notorio que no puede ofrecer ninguna garantía de verdad. También padece una incapacidad para captar las inferencias basadas en el conocimiento que son necesarias para la inteligencia, porque está unida de manera inextricable

a los datos y a la frecuencia de los fenómenos en esos datos. Eso conduce a un problema de cola larga e invoca el fantasma muy real de la improbabilidad y las excepciones. Además, los sistemas inductivos son frágiles, carecen de robustez y no adquieren una comprensión genuina a partir de los datos por sí solos. La inducción no es la senda que llevará a la inteligencia general.

Y tercero, el pensamiento inteligente implica un conocimiento capaz de ir más allá de la observación directa, pero la manera en que adquirimos ese conocimiento es un misterio, igual que la manera en que aplicamos el conocimiento adecuado en el momento oportuno a un problema. Ni la deducción ni la inducción sirven para arrojar luz sobre ese misterio central de la inteligencia humana. La inferencia abductiva que Peirce propuso tanto tiempo atrás sí lo hace, pero no sabemos cómo programarla.

Por tanto, pese a las recientes proclamas que apuntan a lo contrario, no nos encontramos en un camino que vaya a conducirnos a la inteligencia artificial general —al menos, no de momento—. Continuamos enfrascados en la búsqueda de una teoría fundamental.

Capítulo 13

Inferencia y lenguaje 1

Para tratar con Eugene Goostman, tienes que mandarle mensajes de texto. No admite llamadas telefónicas, ni citas para ir a comer. Mándale un mensaje y te contará que es un chico de trece años de la ciudad ucraniana de Odessa. Como los de tantos adolescentes, sus mensajes son frívolos, evasivos, arrogantes; Eugene tiende a jugar al despiste y a irse por las ramas. Se muestra grosero, acto seguido juguetón. Bromea. Lo que no te contará es que en realidad es un programa de ordenador, un bot conversacional diseñado por investigadores rusos para convencer a las personas de que está hecho de carne y hueso.

Se supone que Goostman hizo historia el 7 de junio de 2014, al pasar el test de Turing sesenta años después de la muerte de este. En un muy publicitado evento que contó con la universidad de Reading, Inglaterra, como anfitriona y que se desarrolló en la Royal Society de Londres, el bot conversacional Goostman convenció de que era humano al 33 % de los jueces tras intercambiar mensajes de texto con ellos durante cinco minutos.

Aunque no se tratara de un verdadero test de Turing, el evento llegó a los principales medios de noticias —lo cubrieron blogs y organizaciones de prensa de todo el mundo—. Dos tercios de los jueces no cayeron en los trucos de Goostman, y la prueba se interrumpió a los cinco minutos de comenzar. Aun así, como era de esperar, la prensa desplazada al lugar se mostró eufórica ante la llegada de aquel «momento tan esperado»; el desafío de Turing había sido superado y aquello iba a marcar el comienzo de la IA. The Independent aclamó a Goostman como una «consecución», y añadió que el programa era un «superordenador».[1] Time proclamó que «La Era de los Robots ya está aquí». La BBC lo definió como una «novedad mundial» —lo cual fue técnicamente acertado—, mientras que la popular bitácora de tecnología Gizmodo informaba a sus lectores de que «Esto es importante». Sin duda, la reacción de la prensa dio fe de la fascinación que Alan Turing seguía provocando. También representó un reconocimiento a la trascendencia que tendría superar realmente su test. Imagina, si puedes, la posibilidad de mantener una charla real con tu ordenador. El éxito

de Goostman hizo que ese sueño sempiterno de la ciencia ficción se balanceara delante de nuestros ojos.[2]

EUGENE GOOSTMAN ES UN FRAUDE

Solo hubo un problema: Goostman fue un fraude. Poco después del espectáculo de Londres, que le valió al equipo ruso los cien mil dólares del premio Loebner por haber superado el test de Turing, la comunidad de científicos y analistas comenzó a quejarse. Para comenzar, Goostman ni siquiera contestó a las preguntas que le hicieron. Por lo general cambiaba de tema o soltaba un sarcasmo. Gary Marcus, que por entonces trabajaba en la universidad de Nueva York, escribió en el New Yorker que la actuación de Goostman había sido poco más que un «truco de magia». Sus palabras replicaron las de Hector Levesque, quien también había desdeñado ese tipo de actuaciones como «trucos baratos». Tal y como explicó Marcus, Goostman utilizó una táctica, la de desviar la atención, que genera una ilusión de inteligencia sin necesitar que esta exista:

Marcus: —¿Sueles leer el New Yorker?

Goostman: —Leo montones de libros..., tantos que ni siquiera sé cuáles son.[3]

Estos diálogos son como los que la gente mantenía en la década de 1960 con el programa de chat ELIZA, que imitaba a un psicoanalista rogeriano:

Paciente: —Bueno, estoy teniendo problemas con mi marido.

ELIZA: —Cuéntame ás cosas sobre tu marido.

Aunque cabe reconocerles la gracias, los dos programas fracasan a la hora de obtener cualquier tipo de avance en los desafíos muy reales sobre el lenguaje natural a los que se enfrenta la IA. De hecho, Goostman (y ELIZA) tienen éxito gracias a que evitan la comprensión del lenguaje, y por ese motivo resultan irrelevantes para un examen conversacional serio.

De manera quizá comprensible, a raíz de la conmoción provocada por Goostman, críticos y comentaristas optaron por denunciar el propio test de Turing acusándolo de ser un hito carente de legitimidad. Gizmodo se apresuró a publicar un artículo donde afirmaba que el test de Turing era «una chorrada». Wired se preguntó si el objetivo de mantener una charla al lado del fuego con un ordenador debía ser de verdad el punto de referencia de una IA hecha y derecha.[4]

Si tenemos en cuenta la debacle protagonizada por Goostman, los críticos llevan razón en algo. Hacer trampas a base de espetar respuestas sarcásticas, sin albergar ninguna comprensión real, a duras penas puede medir la capacidad lingüística de la manera en que Turing pretendía. Hacer trampas en el test de Turing recurriendo a ardides y trucos baratos expone un punto débil que tenemos: bajar el listón para detectar la inteligencia cuando la expectativa misma de un diálogo conversacional es en sí baja.[5] Podemos seguirles el juego a los chavales ucranianos de trece años, pensando que no saben tanto como nosotros (ni les importa) sobre los diálogos adultos. Asimismo, podemos asumir durante una de las reuniones de psicoanálisis con ELIZA que la interacción terapéutica incluye un esfuerzo deliberado por hacer que hablemos —lo que en consecuencia excusa al terapeuta de pensar por nosotros—. En este tipo de situaciones, nuestras expectativas encajan en un contexto social que comienza por descartar la evaluación de la inteligencia de quienes responden. Así que no es de extrañar que la mayoría de los investigadores en IA hayan abandonado el desafío del test de Turing. Un comentario desdeñoso por parte de Stuart Russell —«los investigadores en IA ya establecidos no están dedicando casi ningún esfuerzo a superar el test de Turing»— refleja esa frustración ante el frenesí de los medios por las actuaciones con trucos de magia.[6] Aceptarlas es una señal de debilidad para la disciplina.

Pero la desestimación es completamente innecesaria. Para comenzar, un test de Turing honesto sería el punto culminante de la comprensión del lenguaje. Tal y como ha señalado Ray Kurzweil, una inteligencia alienígena quizá no entienda una conversación en inglés, pero toda inteligencia que pase un test de Turing legítimo tiene que ser inteligente. «La cuestión clave es la conversación —

afirma—. A fin de superar el test, tienes que ser inteligente».[7] Kurzweil sugiere que los concursos del futuro se limitarán a permitir tests más largos, para asegurarse de que los trucos baratos se vean filtrados por el diálogo continuado.

Se trata solo de una sugerencia... Quizá, para progresar, tengamos que excluir la contribución de los trucos baratos añadiéndole una regla al juego: el requisito de que los participantes tengan que responder a las preguntas de manera directa, como si se encontraran delante del juez y hubieran jurado «decir la verdad y nada más que la verdad». Las expectativas de un juzgado sin duda descartarían actuaciones parecidas a la de ELIZA. Imagínate contestando a una pregunta del fiscal o del juez con un: «Cuénteme usted mismo más acerca de esa noche. ¿Cómo se sintió?». O podríamos indicar a todos los participantes que deben comportarse como si se encontraran en una entrevista de trabajo para el puesto de entender el inglés conversacional —¡no sería un mal examen para los asistentes personales activados por voz del futuro!—. En tales casos, los trucos representarían una violación inmediata de las reglas propias de la prueba. Goostman estaría hundido.

Los lingüistas informáticos y los investigadores de IA han sabido desde el principio que participar en un diálogo abierto presenta mayor dificultad formal que interpretar un monólogo, como la comprensión de un artículo de periódico. Una forma más de preservar la intuición de Turing según la cual la capacidad de entender el lenguaje natural es una prueba adecuada de la inteligencia de nivel humano consistiría en simplificar el test original y aceptar solo monólogos. Podríamos llevarlo a cabo en el contexto de una sesión de preguntas y respuestas, como con el test original. Piensa en esta simplificación del test, a la que llamaremos el «monólogo del test de Turing». En el monólogo del test de Turing, el juez se limita a pegar un artículo de prensa u otro texto, y a continuación hace preguntas que requieren la comprensión de lo que se dice. El participante debe contestar a estas preguntas de manera correcta. (Se acabaron los trucos.) Por ejemplo, un juez podría pegar el artículo de Associated Press de «¡Los tacos o la vida!» y preguntarle al respondedor si la historia es graciosa o no, y por qué. Al superar esa prueba se alcanzaría, en sentido estricto, un subconjunto lógico de un test de Turing completamente abierto, así que sería totalmente justo utilizarla —de hecho, le concedería una ventaja a la máquina, que posiblemente no sepa comprender del todo cómo lidiar con los «fenómenos pragmáticos» en un diálogo de ida y vuelta—. Luego hablaremos más sobre este tema.

Por desgracia, el monólogo del test de Turing ya nos ha dado pruebas de que los sistemas de IA andan perdidos sin remedio. Gary Marcus y Ernest Davis señalan que la comprensión lectora de las máquinas es lamentable. Los medios celebraron con fuerza a Microsoft y Alibaba por haber mejorado la puntuación estándar en un examen de lectura conocido como Conjunto de datos de comprensión lectora de la universidad de Stanford (SQuAD, en sus siglas inglesas), pero solo cuando en el texto aparecían las respuestas completas. Se trató, por tanto, de una tarea simplificada consistente en «subrayar» las respuestas que se ofrecían de manera explícita, y que venían señaladas por unas preguntas que apuntaban hacia ellas con claridad.[8] Marcus y Davis ponen de relieve actuaciones vergonzosas en preguntas en apariencia simples, como la que preguntaba solo por el nombre del quarterback al que se mencionaba en un recorte sobre la Super Bowl. El fracaso de la comprensión en esos casos resulta evidente. Así que, incluso cuando eliminamos los trucos, la comprensión del lenguaje por parte de la IA se encuentra en problemas. El test de Turing continúa siendo una forma de evaluación legítima que, como mucho, sitúa el listón a demasiada altura.

Cuanto más seria es la mirada que dedicamos a los requisitos de la comprensión del lenguaje, más abrumadora se vuelve la idea de superar siquiera el monólogo del test de Turing. Hector Levesque concibió una versión muy simplificada del test y bautizó sus preguntas como esquemas de Winograd (en honor del pionero de la IA Terry Winograd, que trabajó en la comprensión del lenguaje natural). Los esquemas de Winograd requieren que se conteste a preguntas de opción múltiple sobre el significado de frases simples en inglés. Es algo que se encuentra a años luz del test de Turing. Y, sin embargo, los investigadores de IA están a años luz de dominarlos.

EL CURIOSO CASO DE LOS ESQUEMAS DE WINOGRAD

Hector Levesque es uno de los escasos científicos de IA que hoy en día siguen centrándose en la representación del conocimiento y la razón. De manera admirable, Levesque pretende imbuir a los programas contemporáneos de IA con algo más que técnicas estadísticas para el análisis de los macrodatos: quiere dotarlos de sentido común.

Levesque propuso una versión simplificada del test de Turing que resulta mucho más sencilla que el original, abierto e ilimitado, pero que, y esto es importante, continúa sumiendo en la frustración a todos los enfoques automáticos para la comprensión del lenguaje que conocemos. En 2013, Levesque presentó el artículo «On Our Best Behavior»[«Para portarnos como nunca»] en la Conferencia Internacional Conjunta sobre Inteligencia Artificial [Internet Joint Center for Artificial Intelligence, o IJCAI], y este no tardó en ser reconocido como una llamada a las armas en pos de una IA genuina.[9] Inspirándose en el test completo de Turing, Levesque sugirió que propongamos a las máquinas preguntas que requieran una comprensión más profunda de lo que se está diciendo. Las preguntas del test son frases sencillas, no conversaciones enteras. Por ejemplo: «¿Podría un cocodrilo participar en una carrera de obstáculos?». Levesque escogió deliberadamente preguntas de ese estilo porque las personas que no sean expertas pero dispongan de un sentido común normal podrán responderlas con corrección (no, un cocodrilo no podría participar en una carrera de obstáculos), pero los trucos populares, como usar un motor de búsqueda para dar con la respuesta, no funcionarán. Puesto que no existen (asumimos) páginas web que traten la participación de los cocodrilos en las carreras de obstáculos, no habrá manera de eludir la necesidad de una comprensión. Ante esos ejemplos, los sistemas de IA se encallan, pero a las personas se les ocurre la respuesta casi de manera inmediata.

El esquema de Winograd es un ejercicio de opción múltiple que anula la posibilidad de que la máquina pueda desviar la atención, recurrir al sarcasmo, bromear o aparentar malhumor para engañar a los jueces humanos en cualquier situación en la que una respuesta directa revelaría su falta de comprensión. Los esquemas se basan en un rasgo común del lenguaje natural, tal y como deja a las claras la pregunta original de Winograd que les sirve de inspiración:

Los concejales del pueblo se negaron a conceder un permiso a los manifestantes rabiosos porque temían que hubiera violencia. ¿Quién temía esa violencia?

A) Los concejales del pueblo

B) Los manifestantes rabiosos

Fíjate en que el sujeto elíptico de «temían» es el pronombre plural «ellos», que puede referirse tanto a los concejales como a los manifestantes. En otras palabras, es ambiguo porque ambas respuestas son posibles sin quebrantar las reglas de la gramática. Sin embargo, solo una de ellas tiene sentido. Las personas ofrecen la respuesta correcta a estas preguntas sin el menor esfuerzo y con un acierto cercano al 100 %. Los sistemas de IA, no. Su rendimiento en los esquemas de Winograd no mejora demasiado el de las respuestas que se dan al azar.[10]

Junto con otros investigadores de IA, Levesque exhumó el desafío de Winograd en 2012, cuando el big data y el aprendizaje automático dominaban con claridad los enfoques en busca de la IA (sigue siendo así). Reunió un conjunto de preguntas de opción múltiple que explotaban el rasgo común y ubicuo de la ambigüedad en el lenguaje natural. Llamó a ese desafío el problema de la desambiguación de los pronombres. Con él, captó la inspiración tras el test de Turing, pero de manera simplificada: la comprensión de un lenguaje natural común y corriente, como el inglés o el francés (o el que tú quieras) requiere de una inteligencia general. En especial, Levesque pensó que los sistemas de IA necesitarían conocer lo que las palabras del lenguaje significan en realidad para superar el test. He aquí otro ejemplo de esquema de Winograd:

Joan se aseguró de agradecerle a Susan toda la ayuda que le había ofrecido. ¿Quién ofreció ayuda a quién?

A) Joan

B) Susan

A fin de proteger los esquemas del problema de los trucos baratos que se emplearon para engañar a los jueces del test de Turing, Levesque añadió un giro: dos palabras designadas como especiales que «le dan la vuelta» a la respuesta, dejando el resto de la pregunta inalterada. En este ejemplo, las palabras especiales son ofrecido y solicitado. Al intercambiar esas palabras especialmente designadas se genera otra pregunta:

Joan se aseguró de agradecerle a Susan toda la ayuda que le había solicitado. ¿Quién solicitó ayuda a quién?

A) Joan

B) Susan

He aquí otro esquema, que emplea las palabras especiales jugadores de golf y perros:

Sam intentó pintar un cuadro de pastores y ovejas, pero acabaron pareciendo jugadores de golf. ¿Quiénes parecían jugadores de golf?

A) Los pastores

B) Las ovejas

Los esquemas de Winograd resultan muy simples en comparación con el test original y conversacional de Turing, pero, al proponer preguntas de opción múltiple que obligan a resolver la referencia pronominal (la desambiguación pronominal), capturan lo que investigadores anteriores llamaron «holismo del sentido común» —la idea de que el lenguaje natural no se puede comprender solo a través de la disección de las frases, sino que requiere de una comprensión general—. Así, los esquemas de Winograd son preguntas sencillas, pero por lo general poco comunes, que tienen perfecto sentido para aquellos lectores con un conocimiento normal del mundo. Por tomar el ejemplo de Marcus, con un conocimiento básico sobre los cocodrilos y las vallas, resulta evidente que las cortas patitas de los primeros impiden que puedan competir saltando esas vallas. ¿Por qué esto representa un problema tan difícil para la IA?

En parte, la dificultad de los esquemas se debe a que los dos referentes —la elección de nombres y sintagmas nominales como «cocodrilo»y «carrera de obstáculos» rara vez aparecen juntos (si es que llegan a hacerlo) en páginas web y otros textos. Puesto que los enfoques de IA basados en datos dependen del

análisis estadístico de numerosos ejemplos, preguntas extrañas como las que encontramos en los esquemas de Winograd constituyen un desafío notable —todo ello en una sola frase—. De hecho, los esquemas de Winograd son bastante inmunes a trucos como el del conteo de páginas web. Están hechos «a prueba de Google», en palabras de Levesque. Pero la necesidad de contar con un conocimiento cotidiano para interpretarlos sigue siendo una razón de mayor peso para que a los ordenadores no se les den bien. Cambiar el sujeto de Marcus, de cocodrilos a gacelas, haría cambiar también la respuesta (las gacelas pueden saltar obstáculos sin problemas), pero la pregunta sigue siendo extraña, así que resulta infrecuente para la red. El aprendizaje automático y el big data no sirven de ayuda. Los sistemas de IA no pueden buscar la respuesta.

Como cabía esperar, los investigadores de IA han desarrollado nuevos trucos para afrontar los esquemas de Winograd. Puesto que hoy en día, en la web, existe tanto contenido disponible para el análisis, en algunos casos aún hay patrones de frecuencia que se pueden explotar. Por ejemplo, para responder a preguntas sobre si x es más alto que y, los investigadores pueden imitar el conocimiento real realizando búsquedas en la red con el patrón «x es más alto que y». Si resulta que las personas, edificios o lo que sea que representen x e y se encuentran en depósitos como Wikipedia, los datos estarán disponibles y se podrá usar la matemática para responder a la pregunta —todo ello procesando campos de información y otras fuentes de datos en la red—. Pero, una vez más, Levesque anticipó esa técnica y sugirió ejemplos con nombres comunes (gacelas, cocodrilos, bolígrafos, papel, bolas de bolos, ovejas, etc.) que no aparecen en las bases de datos verificadas ni en las enciclopedias online. Eso obstaculiza el truco de buscar en la red a la vez que subraya la cualidad cotidiana, de sentido común, del conocimiento requerido —lo cual preserva la voluntad original de Turing de utilizar las conversaciones básicas a las que se da todo el mundo en la vida diaria—.

Los esquemas de Winograd se protegen de los trucos relacionados con los motores de búsqueda de otra manera importante. La relación entre x e y también se puede modificar, igual que las cosas sobre las que se habla (cocodrilos frente a gacelas). Se trata de otro obstáculo al uso de trucos con los datos a la vez que se mantiene el carácter simple de las preguntas. Piensa, primero, en un esquema que requiera un conocimiento sobre los tamaños relativos de los objetos comunes.

El trofeo no cabía en el maletín marrón porque era demasiado pequeño. ¿Qué era demasiado pequeño?

A) El trofeo

B) El maletín marrón

En realidad, para este esquema existe un truco —una técnica—. Tal y como señala Levesque con un lenguaje casi computacional, podemos diseccionar la pregunta en una relación (llámemosla R) y una única propiedad (llámemosla P).

R = no cabe en

P = es demasiado pequeño

A continuación, usamos el «big data», dice, y «buscamos en todo el texto de la red para determinar cuál es el patrón más común: “x no cabe en y” + “x es demasiado pequeño” frente a “x no cabe en y” + “y es demasiado pequeño”».[11] Ante el montón de ejemplos de frases que nos devuelve la búsqueda podría aflorar un patrón según el cual, por ejemplo, el segundo objeto mencionado sea con mayor frecuencia más pequeño que el primero. Imagínate una aportación en una red social donde alguien se queja de que está guardando sus cosas y no deja de decir que esto y aquello no le cabe en la mochila porque esta es demasiado pequeña. El mero conteo de los resultados de una y otra búsqueda para ambos patrones puede conducir a una respuesta estadística. Y, de ser esta correcta, el sistema podría responder a un esquema de Winograd sin disponer de conocimiento.

Por desgracia, ese método es bastante superficial. Se encuentra indefenso ante las modificaciones más pequeñas, que alteren el sentido de la pregunta y que por consiguiente hagan variar la preferencia de la respuesta. El truco original de la «relación R» falla, por ejemplo, con esta pregunta:

El trofeo no cabía en el maletín marrón pese a que era muy pequeño. ¿Qué era muy pequeño?

A) El trofeo

B) El maletín marrón

Ahora es el trofeo, y no el maletín, lo que es pequeño. De modo que las especificaciones de R y de P generarán una respuesta incorrecta. La pregunta modificada es con toda probabilidad menos frecuente que su versión original, así que el big data es un estorbo, no una ayuda. El problema es el siguiente: incluso con preguntas sencillas, de una sola frase, el significado de las palabras —«cocodrilos, obstáculos»— y el significado de las relaciones entre las cosas —«ser más pequeño que»— estropean las técnicas que dependen de datos y frecuencias sin llegar a la comprensión. Los esquemas de Winograd son una ventana hacia una dificultad muy superior a la del test de Turing.

Durante la última década, el aprendizaje automático y los macrodatos han realizado avances importantes en diferentes problemas. Por lo general, no obstante, esos éxitos se han debido a soluciones alternativas que esquivaban un conocimiento y una comprensión reales. Aquí se da una confusión recurrente, en especial con los sistemas automatizados de traducción del lenguaje, que parecen requerir algún tipo de comprensión de ese lenguaje. Fíjate en Google Translate, que a menudo se promociona como la historia de un éxito arrollador y la demostración de que la IA está dominando cada vez con más rapidez el lenguaje natural. Pero es que la disponibilidad más o menos reciente de amplios volúmenes de textos traducidos en la red facilita los enfoques de datos intensivos, que ofrecen resultados «razonablemente buenos» explotando el mapeo de frases y palabras en textos traducidos a lenguajes diferentes. La mayor parte de los mapeos se encuentran en los datos, así que las estrategias inductivas funcionan. No hace falta ninguna comprensión.

Sin ir más lejos, los documentos parlamentarios oficiales de Canadá se traducen palabra por palabra del inglés al francés; en casos así, métodos como el del aprendizaje profundo se bastan para «aprender» los mapeos entre ambos lenguajes. No hace falta ninguna comprensión del inglés o del francés. Por ejemplo, si escribes «John conoció a Mary en el café» en Google Translate y

solicitas su traducción al francés, obtendrás «John a rencontré Mary au café». Se trata de una traducción perfectamente aceptable. Sin embargo, otros ejemplos fracasan. Y, por lo general, esos fallos implican algún tipo de ambigüedad, con fenómenos referenciales como los pronombres, o que en el contexto haya integradas palabras polisémicas (de varios significados). (Si quieres experimentar en Google Translate con frases que contengan alguna ambigüedad, prueba con The box is in the pen*e o con «Me encantó el río. Me fui hasta el banco»).

Para mayor confusión, en los años sesenta el trabajo con la denominada traducción automática de alta calidad completamente automatizada se inició con los enfoques estadísticos, si bien aquellos eran más simples. No funcionaron demasiado bien, y apóstatas como Yehoshua Bar-Hillel no tardaron en llegar a la conclusión de que no había esperanza para la traducción automática, porque parecía necesitar un conocimiento y una apreciación del contexto —motivos que más tarde inspiraron a Winograd, y a continuación a Levesque, para formular esquemas basados en el conocimiento—. Pero Bar-Hillel no previó la contribución que los macrodatos iban a realizar algunas décadas después. Los enfoques estadísticos comenzaron a revelarse prometedores en la década de 1980, con el trabajo de los investigadores de IBM, y gigantes de la red como Google utilizaron técnicas similares (y, ahora, el aprendizaje automático) para ofrecer servicios de traducción decentes en términos estadísticos. La traducción automática de alta calidad completamente automatizada había cerrado el círculo.

Pero el escepticismo original de Bar-Hillel sigue resultando pertinente; las traducciones que requieren un contexto y unos conocimientos siguen siendo una caja negra para los métodos contemporáneos, con su big data y demás. Por ejemplo (e irónicamente), en octubre de 2020, Google Translate sigue equivocándose con el ejemplo que Bar-Hillel puso en los años sesenta. Bar-Hillel preguntó cómo programar una máquina para que tradujera «The box is in the pen» de manera correcta. Aquí, «pen» resulta ambiguo. Puede significar un elemento de escritura, o un recinto cerrado para contener a los animales. Usando Google Translate, «The box is in the pen» conduce en francés a «La boîte est dans le stylo», donde «stylo» significa instrumento de escritura, que no es la traducción preferida (porque en general las cajas son más grandes que los bolígrafos). En otras palabras, la traducción «razonablemente buena» depende de nuestros propios requisitos. Es posible que Google Translate reciba la mayoría de las veces frases menos contextualizadas, pero cometerá errores —y esos errores podrán ser de importancia, dependiendo de la persona que lo utilice—.

En realidad, servicios como Google Translate enfatizan el problema de cola larga de los enfoques estadísticos o inductivos, que van empeorando a medida que los ejemplos y las interpretaciones se vuelven menos probables. Esta es otra manera de decir que la probabilidad no equivale a una comprensión genuina. Ni siquiera se encuentran en el mismo espacio conceptual. Aquí, de nuevo, la suposición de la frecuencia representa una limitación real para acceder a la inteligencia artificial general.

Los resultados «razonablemente buenos» de la IA contemporánea son en sí una especie de truco, una ilusión que enmascara la necesidad de una comprensión cuando se lee o se conversa. En las tareas de reconocimiento de imágenes, quizá se les pueda perdonar a los sistemas que clasifiquen ciertas imágenes de manera incorrecta (pero no si el sistema es un coche sin conductor). Sin embargo, en las pruebas de lenguaje como los esquemas de Winograd, una tasa de error del 20 % (es decir, una tasa de acierto del 80 %) implica que dos de cada diez ejemplos son opacos para el sistema. Y la tasa de exactitud en los tests con esquemas de Winograd es mucho peor: apenas supera la de las respuestas ofrecidas al azar. Tal y como dejan a las claras las preguntas de los esquemas, los errores resultan menos excusables cuando se dan en frases simples, sin la menor dificultad inherente para las personas.

La cuestión es que la exactitud misma es contextual y, en pruebas que exponen la ausencia de cualquier comprensión, obtener seis respuestas correctas de cada diez (como con los sistemas más avanzados) no representa ningún progreso. Es la demostración de una idiocia disfrazada. Igual que los míticos androides del thriller de ciencia ficción Blade Runner, una prueba que busque emociones mentales (o reales) acabará por desenmascarar a las máquinas como impostores programados. Lo que pasa es que los esquemas de Winograd son cortos y bastante simples. No hace falta más que una pregunta que requiere de un pequeño conocimiento básico de lo que se está diciendo.

El «problema de la desambiguación de los pronombres» de Levesque es un simple paso en el vastísimo campo de minas que los sistemas automáticos tendrán que atravesar si quieren tener alguna esperanza de resolver las conversaciones cotidianas. La resolución de las referencias pronominales es apenas uno más de la plétora de problemas basados en el conocimiento que deben resolverse de cara a alcanzar una comprensión genuina del lenguaje. Y los esquemas de Winograd también representan una simplificación enorme en otros sentidos. Junto con el límite de una sola frase que tiene cada esquema, la prueba

ignora la mayoría de los fenómenos pragmáticos que surgen en un diálogo conversacional. La pragmática es el infierno de la automatización. Para ver el porqué, nos iremos de paseo por el ámbito del procesamiento del lenguaje natural, o la comprensión.

Capítulo 14

Inferencia y lenguaje 2

El lenguaje natural es un animal complejo. Sus ingredientes son unos símbolos, las letras y la puntuación, que forman palabras: p-a-l-a-b-r-a-s. Las palabras son fragmentos de significado. Hacen referencia a todo tipo de cosas: objetos como una «cuchilla», sensaciones como «placentero», juicios morales como «malo», situaciones o acontecimientos como «fiesta» o «elecciones», y grupos abstractos como la «economía» o la «granja». Aun teniendo significado, las palabras pueden resultar ambiguas: «granja» puede referirse a la localización física de una granja, o al negocio de la granja, o al lugar en el que vive la familia. El sentido exacto de una palabra tan simple como «granja» dependerá de la verdadera voluntad del hablante, que es contextual.

Comprender el contexto amplía el alcance de la interpretación del lenguaje a las palabras, frases y párrafos circundantes, e incluso a textos enteros. Se debe resolver la ambigüedad en fenómenos como el de los pronombres y la indicidad: dónde y cuándo («a la hora acordada», cuando el contexto deja claro que se refiere a la medianoche). Por último, tenemos lo que los lingüistas llaman «pragmática»: el contexto que incluye a la persona que realiza las afirmaciones, sus propósitos, intereses y demás. El sarcasmo, la ironía y otros aspectos de la comunicación entran en juego. El lenguaje comienza con las letras y las palabras, y termina con preguntas sobre la mente y el significado.

UN PASEO POR LA COMPRENSIÓN DEL LENGUAJE

La comprensión del lenguaje (pongamos que la capacidad de comprender un tuit, un blog o Guerra y paz) es una especie de pirámide invertida, como la de las viejas ilustraciones de la pirámide alimentaria que se usaban para mostrar una dieta equilibrada, pero en este caso del revés. A medida que asciendes por esa

pirámide (invertida) vas encontrando más significado, y los tratamientos informáticos del lenguaje se topan cada vez con mayores dificultades. Abajo del todo, en la punta, tenemos la ortografía, las letras y la puntuación que, combinadas, forman palabras y frases. Siguen los rasgos léxicos y morfológicos como las palabras, sus prefijos, sufijos, categorías gramaticales, etc. Para formar las oraciones necesitamos palabras que se combinen en forma de frases, como el sintagma nominal «el león feroz», que contiene un sujeto e información identificativa y calificativa; a continuación hay que añadirle una acción, como en «el león feroz devora», y también un objeto si la acción (y el verbo) son transitivos, como en «el león feroz devora la carne». Las oraciones forman párrafos, que expresan un tema y le dan cuerpo, y la unión de múltiples párrafos conforma una historia o narración. La historia (de uno o más párrafos) se llama «discurso» o simplemente «texto». Un ejemplo de texto amplio sería el de un libro, que cuenta con un género, como ficción o no ficción, etc. La transcripción de una conversación también es un texto. Así pues, el test de Turing descansa en lo alto de la pirámide invertida del lenguaje, donde este se amplía y expande en forma de diálogo completo, un texto conversacional.

En cierto sentido, se puede ver la pirámide invertida del lenguaje de manera puramente sintáctica, como la construcción de un texto según unas reglas que especifican cómo combinar caracteres en palabras, palabras en oraciones y oraciones en párrafos. La sintaxis se puede procesar y analizar por ordenador. Pero la pirámide del lenguaje lleva una doble vida: también trata sobre el significado. Los lenguajes naturales se sirven de símbolos y reglas para transmitir un significado, cosa que no se puede ignorar si queremos decir algo útil usando el lenguaje. El lenguaje es un trayecto hacia el significado. Cualquier libro infantil sigue el mismo trayecto que una novela rusa. Todo es la expresión de un significado.

En la lingüística y otros ámbitos afines, la semántica es el estudio del significado de las cosas. «Semántica» significa «significado», y el análisis semántico es aproximadamente lo que hace falta para realizar un test de opción múltiple con los esquemas de Winograd. Se requiere un conocimiento y unas convicciones, no solo sintaxis e información acerca de los símbolos. Es necesaria una inferencia para aplicar ese conocimiento en su contexto. Por ejemplo, si creo que un cocodrilo puede correr los cien metros vallas, tengo que explicar la manera en que sus patitas han logrado dar los saltos de rigor por encima de los obstáculos, y para hacer eso necesito saber algo sobre la anatomía de los cocodrilos. (Con las personas no hacen falta grandes explicaciones, de hecho: bastará una foto, como

la que cualquier niño de cinco años habrá visto con toda probabilidad.) La respuesta a esa cuestión no se halla en la sintaxis, sino en la semántica.

Los esquemas de Winograd reemplazan la pirámide sintáctica del lenguaje con la semántica, donde las palabras tienen significados que se refieren a las cosas del mundo. Pero, a veces, las palabras de una oración solo tienen sentido en el contexto de un texto entero. La interpretación textual siempre implica consideraciones pragmáticas, que añaden al significado de las palabras inferencias sobre propósitos e intenciones. Una conversación entre dos personas sirve como ejemplo de la pragmática en la comprensión del lenguaje. Si te pido «¿Puedes pasarme la sal?» y contestas «Sí» pero no haces nada, el contexto —mi intención— se te habrá escapado (o estarás bromeando, lo cual sigue siendo pragmática). En la pragmática, lo que se dice mantiene una estrecha dependencia con la manera en que se dice y el motivo por el que se dice.[1] Lo que la gente quiere decir casi nunca es una función literal de lo que dicen palabra por palabra. Ese rasgo del lenguaje cotidiano, que la lingüística estudia como pragmática, es lo que hace que la interpretación del lenguaje resulte difícil para la IA, pero elocuente e interesante —y, por lo general, natural y sencilla— para la gente.

La sintaxis sigue siendo esencial. Cuando pasamos de un lenguaje como el inglés a, pongamos, el chino, también estaremos pasando de los caracteres latinos a los pictogramas. Si alguien no habla ni lee el chino, cualquier intento de comunicación será inútil. Pero, tal y como dicen los lógicos, el conocimiento de la sintaxis es necesario, pero no suficiente. El conocimiento del mundo resulta esencial. Si te cuento que han golpeado a alguien con un bate hecho de poliestireno, tendrá importancia lo que sepas acerca del poliestireno, ya que ese conocimiento determinará lo que pienses sobre mi comentario y tu manera de reaccionar: «Qué tragedia», podrías decir, sarcásticamente. Pero un sistema de IA que asumiera que el poliestireno es una sustancia correcta y dura para golpear con un bate podría alarmarse bastante: «¿Se encuentra bien?». Por lo general, incluso el más sencillo de los diálogos necesita de un conocimiento real y unas inferencias dependientes del contexto, lo que involucra a la semántica (y la pragmática).

Turing ya lo entendió en su exposición original de las posibilidades de la IA. En su ejemplo, se imagina a una persona que pregunta (el interrogador) y a un ordenador (el testigo). El interrogador le ha pedido al testigo que escriba un soneto, y este ha reproducido el Soneto 18 de Shakespeare:

Interrogador: —En En el primer verso de su soneto, que dice: «¿A un día de verano compararte?»,*f ¿no funcionaría igual o incluso mejor hablar de «un día de primavera»?

Testigo: —Se rompería la métrica.

Interrogador: —¿Qué hay de «un día de invierno»? Eso respetaría la métrica.

Testigo: —Sí, pero a nadie le gusta que lo comparen con un día de invierno.

Interrogador: —¿Diría que el señor Pickwick le hace pensar en la Navidad?

Testigo: —En cierto sentido.

Interrogador: —Sin embargo, la Navidad es un día de invierno, y dudo que al señor Pickwick le importara la comparación.

Testigo: —No lo dirá en serio... Al hablar de un día de invierno nos referimos a un día gris y frío, no a una jornada especial como la Navidad.[2]

La primera pregunta del interrogador pone a prueba el conocimiento del ordenador acerca de la redacción de sonetos, un conocimiento especializado cuya ausencia le perdonaríamos a gente completamente culta. El pentámetro yámbico*g excluye el día de «primavera» y el ordenador contesta cortés que la sugerencia del interrogador «rompería la métrica».

Pero «día de invierno» no la rompe, así que ¿por qué no? Aquí, el foco de atención se desplaza hacia la metáfora de comparar el amor de uno con un «día de invierno». El testigo ningunea la sugerencia del interrogador para que use esa opción. Podemos asumir que los días de invierno son cortos y fríos, y que una persona profundamente enamorada no escogería esa imagen. Pero, en realidad, el cambio de verano a invierno no es tan evidente. Los días invernales pueden ser hermosos y nevados, con un sol radiante que ilumine el dulce silencio de los bosques. La idea apunta más bien a que, estando enamorados, si hemos de escoger una estación para describir al amado o amada, los días cálidos, largos y hermosos del verano tienen más sentido.

El interrogador sugiere entonces un día «especial» de invierno, la Navidad. Eso pone a prueba el conocimiento del ordenador sobre la intención o propósito del soneto, que no consiste en comparar a alguien con una festividad, sino con algo encantador y «constante», algo parecido a la experiencia de un delicioso día de verano. Comparar un amor con la Navidad, o con el Cuatro de Julio, implicaría mezclar ideas y emociones que mitigarían y estropearían el sentido del soneto. El testigo, en apariencia consciente de tan absurdo engaño, contraataca poniendo en duda la seriedad del interrogador.

En otras palabras, entender el sentido del soneto implica entender la emoción profunda de quien está enamorado y compara a su amado o amada con un día de verano, algo hermoso y prolongado y encantador, si bien el poeta considera que su amor lo es más, inmortalizado por la profundidad de su pasión y por el carácter atemporal del soneto que ha escrito. El diálogo entre el interrogador y el testigo requiere un conocimiento, sin duda, pero en realidad la mayor parte del conocimiento relevante tiene que ver con las personas —con una persona enamorada y la manera en que expresa ese amor con palabras y sentimientos—. Para entender el soneto tenemos que saber lo que se siente al hablar sobre la persona a la que amas. El interrogatorio trata de estados mentales y emociones. Eso es la pragmática, en su sentido más amplio e intencionado.

Podríamos decir que la poesía es intrínsecamente pragmática. Pero la prosa, menos efusiva, no deja de mostrar la ubicuidad de los fenómenos pragmáticos en el lenguaje, casi en cada frase «normal y corriente». Es ineludible. La «deixis», por ejemplo, se refiere a palabras como los pronombres «yo» o «él», o el adverbio «aquí», que no se pueden entender o desambiguar sin conocer el contexto. «Deixis» significa «señalar», y eso es lo que el lenguaje hace a menudo: «Anoche, ella vio un gorrión por aquí» señala a quien viera el gorrión (deixis personal), y el oyente lo entiende por el contexto. Sucede lo mismo con «aquí», que se refiere a alguna localización en el contexto (deixis de lugar), y con «anoche», que debe de referirse a la noche anterior al comentario (deixis de tiempo). Aquel sistema que busque mantener una conversación sin recurrir a trucos preparados debe estar al tanto de todo esto.

Como mínimo, la deixis es solo un problema entre muchos otros. El lenguaje está plagado de sutilezas contextuales que reclaman una comprensión profunda de la intencionalidad. Si menciono que anoche estuve tocando a Mozart, el comentario se ajusta con naturalidad para que tenga sentido: anoche estuve tocando música compuesta por Mozart. (No estuve «tocando» a Mozart como

persona, pongamos que manoseándolo o utilizándolo como si fuera un instrumento musical.) Por lo general, no nos preocupamos en explicitar esos ajustes porque los fenómenos pragmáticos del lenguaje forman parte integral de la comunicación cotidiana. Inferimos el significado que se pretende transmitir no a través de un montón de ejemplos, sino sopesando el mensaje.

Los lingüistas (y el resto de nosotros) saben que, a menudo, la gente quiere decir mucho más de lo que dice. Esto se revela en el lenguaje de diferentes maneras. Las conversaciones están llenas de elipsis, por ejemplo, en las que nombres, verbos y otras categorías gramaticales, y frases también, quedan excluidas por completo: «Ayer fui al centro comercial, y Shana ha ido hoy». Disponemos de un conocimiento previo y específico de personas, lugares y cosas, y lo usamos para acortar la comunicación entre nosotros y para que esta resulte más natural. «¿Me prestas el libro?» asume que tienes un libro, y también que sabes a cuál me refiero. Si me informas de que «Charles dejó de estudiar matemáticas después de ganar el premio de poesía», presupondré que conoces a Charles y que Charles estudiaba matemáticas, etc. Podría decirse que el lenguaje está ocluido en parte: excluimos detalles y suposiciones; contamos con las mentes ajenas para que infieran y asuman lo que queremos decir.

La anáfora, que ya hemos mencionado antes, significa «repetición». En «El barco abandonó el puerto en mayo. Roger estaba en él», el pronombre «él» es una repetición de «barco».[3] Como hemos visto, la anáfora (como su prima, la catáfora, que anticipa algo que va a venir) se introduce en el lenguaje con referentes, por lo general pronombres del tipo «él», «ella», «ellos», «lo», «los», etc. Los pronombres no son especialmente raros, y las anáforas en general están por todas partes. Son ejemplo de un lenguaje que se señala a sí mismo, y que a la vez señala hacia el mundo exterior.

Las anáforas también interactúan con otras inferencias pragmáticas para transmitir significado, como en «Nos encantó el cuarteto de anoche. Todo ello fue encantador», donde el pronombre anafórico «ello» se desplaza con sutileza hasta abarcar la velada al completo, y no solo el cuarteto. La anáfora en sí es un caso especial —pero común— de referencia en el lenguaje. Un tipo popular de esta en los tratamientos informáticos del lenguaje lleva el nombre de correferencia; en ella, dos menciones (palabras o frases que mencionan algo) en un discurso o texto se refieren al mismo objeto o situación del mundo. En un ejemplo anterior, «el barco» y «él» se correfieren al barco que abandonó el puerto en mayo. (Si Roger hubiera estado «junto a él» en vez de «en él»,

podríamos imaginar situaciones en las que se encontrara en el puerto: «El barco estuvo a punto de ahogarle al pasar junto a él».) En efecto, los esquemas de Winograd simplifican el problema de la correferencia general, informática, al reducirla a una sola oración (en vez de, pongamos, a un texto completo) que siempre contiene dos antecedentes posibles: nombres o sintagmas nominales en la oración, donde uno, solo uno, correfiere con el pronombre:

Habían colocado el saco de patatas debajo del saco de harina, así que tuvieron que moverlo antes. ¿Qué hubo que mover antes?

A) El saco de patatas

B) El saco de harina

El esquema plantea un problema de correferencia; el pronombre enclítico «-lo» se refiere al mismo objeto del mundo que el sintagma nominal «el saco de harina» (ambos se refieren a un saco de harina, el que está encima del saco de patatas que se describe en la oración). Pero los esquemas lo simplifican, y mucho. Para comenzar, eliminan la referencia abierta: «Más tarde, ese mismo año, las importaciones cayeron vertiginosamente. A finales de año, la economía al completo tenía problemas». Aquí, «la economía al completo» referencia algo (la economía al completo) aunque sin un adjunto en el texto mismo, como sucede en los esquemas. (En este ejemplo se da una meronimia implícita, una relación parte-todo entre las importaciones y el total de la economía.) Por el contrario, los esquemas de Winograd siempre ofrecen dos opciones a quien ha de interpretarlos, y estas se especifican con claridad.

Las conversaciones, los editoriales de la prensa y la carta a un amigo abocan a una labor interpretativa mucho más complicada. La simplificación del esquema es en parte estructural: puesto que al realizar el test sabemos que solo hay dos respuestas posibles, la estrategia de escoger al azar no funcionará tan mal, se encontrará en torno al 50 % de aciertos. No está mal... ¡y con una inteligencia nula! Pero las personas obtienen resultados del 95 % o más en los esquemas de Winograd, y el desempeño de los mejores sistemas hasta la fecha no es mucho mejor que el de la heurística simple, que escoge siempre la primera opción o lanza suposiciones al azar.[4] Las anáforas son un conocimiento explícito y una

condición inferencial sobre la interpretación —no una prueba de frecuencias en los datos—. Esa conclusión resulta especialmente devastadora porque en cualquier caso el test no esconde ningún truco. Las preguntas están articuladas por completo, y las elecciones son claras. En su mayoría, la gente piensa que son sencillas (aunque quizá un poco raras).

Pensemos en esa prueba de lenguaje mucho más compleja, el monólogo del test de Turing, que hemos descrito antes, donde el sistema de IA debe leer un fragmento de monólogo, como un relato o un artículo de prensa, y responder a las preguntas sobre él que realiza un juez humano. He aquí la historia de uno de esos tests:

Con la esperanza de iniciar una nueva vida, un hombre se trasladó a otra ciudad después de renunciar a su trabajo. Se dirigió en coche hasta el supermercado del lugar para ver si su descuento como empleado seguía siendo válido. A la salida, en la caja automática, descubrió que no era así. Murmuró para sí mismo: «No pienso volver aquí nunca más». Una mujer le oyó y le dirigió una mirada suspicaz. Al salir se topó con su esposa y le contó lo que había sucedido. Repitió: «¡No pienso volver nunca más por aquí!». Ella le contestó con una sonrisa: «Pues yo sí. Está cerca de casa y es barato». Él asintió con la cabeza, sonriente, y dijo: «Me parece bien». Más tarde fueron al parque y él la comparó con un día de verano.

La historia es forzada (me la he inventado), pero no resulta especialmente difícil para la comprensión del lector, y tampoco es que requiera de un conocimiento de sentido común y de la pragmática únicos de cara a interpretarla de manera correcta. Pongamos como ejemplo que Ultra++, una empresa emergente que está en boca de todos (y que seguramente comprará Google), acaba de lanzar la versión demo de DeepRead, un nuevo sistema de IA que utiliza el mejor aprendizaje automático y el mejor big data disponible. El juez humano le hace algunas preguntas sencillas:

(1) ¿El hombre llegó caminando a la nueva ciudad?

(2) ¿Por qué esperaba el hombre comenzar una nueva vida?

(3) ¿El hombre se murió?

(4) ¿Qué crees que conducía, un coche o un carro de golf?

(5) ¿Solía el hombre trabajar en esa tienda o en otra tienda de la misma compañía?

(6) ¿Seguía siendo válido su descuento de empleado?

(7) ¿El hombre murmuró que nunca más volvería a la tienda? ¿O se refería al pueblo? ¿O al espacio frente al autoservicio?

(8) La mujer que le oyó murmurar ¿era su esposa?

(9) ¿Qué quiso comunicarle a su esposa cuando se topó con ella?

(10) ¿Qué quiso decir ella con ese «Pues yo sí»?

(11) La tienda ¿estaba al lado de su casa?

(12) ¿El hombre amaba a su mujer?

(13) ¿Crees que inició una nueva vida, después de todo?

Si aspira a responder a esas preguntas, DeepRead tendrá que resolver diversos problemas complejos de la pirámide del lenguaje que están relacionados con la semántica y la pragmática, y ninguno de ellos aparece de manera estricta en los datos, en el sentido que requeriría una IA datocéntrica.

Comencemos por el conocimiento de sentido común según el cual «trasladarse a otra ciudad» implica hacer las maletas y coger el coche o subirse a un avión para llegar hasta ella. En sentido estricto, la pregunta (2) no se puede responder pero, dado el contexto, la mayoría de los lectores humanos asumirían que esa «nueva vida» está relacionada con el reciente abandono de su trabajo (aunque un lector menos intrépido podría sostener de manera razonable que no tiene ni idea). La pregunta (3) también requiere que se comprenda que el modismo «iniciar una

nueva vida» no significa que hayas de morirte físicamente antes. La pregunta (4) es un ejemplo de implicatura, o lo que una oración o frase implican sin llegar a afirmar de manera directa. Asumimos que el hombre llegó en coche —aunque existe la posibilidad lógica de que llegara en un autito de golf—. En las preguntas restantes, DeepRead debe usar sus conocimientos previos para darse cuenta de que el descuento de empleado que tiene el hombre funcionó en su día, así que se puede presumir que trabajó para la compañía dueña de la tienda en la que se encuentra (aunque en un lugar diferente —se supone que en la ciudad que ha abandonado—); deberá resolver el pronombre «le» sobre el hombre (resolver la anáfora); deberá comprender la deixis de lugar «aquí» en referencia a la tienda (no a la ciudad o al espacio delante de la caja de autoservicio); deberá comprender que «una mujer» no se puede referir a su esposa (otra implicatura); deberá comprender que el hombre informó a su esposa no solo acerca de su intención de no volver por esa tienda, sino de que se sentía frustrado con la empresa —y la lista de problemas de lenguaje va creciendo.

Que su esposa diga «Pues yo sí» es un mensaje incompleto, implica una elipsis, donde la afirmación se acorta a raíz del conocimiento asumido entre el hablante y el oyente. DeepRead debería entender que su esposa volverá a esa tienda, algo que no se dice. Más anáfora y elipsis: ese «Está cerca de casa se refiere a la tienda», y esa cercanía es, junto con el hecho de que sea barata, el motivo por el que la esposa seguirá comprando allí. DeepRead también debería captar que la tienda no está al lado de la casa nueva, ya que el hombre ha llegado a ella en coche. Y, por último, deberá captar que la sonrisa del hombre implica que se liberó de su frustración, algo que remarca el hecho de que más tarde se fuera al parque con su esposa y la comparara con un día de verano. DeepRead debería llegar a la conclusión de que el hombre ama a su esposa y respeta su opinión (lo más probable es que se alegrara de haberse encontrado con ella), y que al compararla con un día de verano estaba expresando su felicidad y su aprecio.

Pero, por desgracia, me he dejado todo tipo de conocimientos necesarios e inferencias. DeepRead debería comprender que el hombre entró en la tienda, y no solo que condujo hasta ella, aunque eso se presente de manera implícita más que como una afirmación. Etc. Ningún sistema de IA existente o predecible puede responder a esas preguntas —y solo estamos hablando del monólogo del test de Turing, sirviéndonos de una historia sencilla—. DeepRead no existe. La mente se adentra en el lenguaje y compone una imagen de lo que está sucediendo, y de su porqué. La pirámide invertida se expande hacia unos problemas que requieren de una imagen del mundo profunda y cargada de

significado, y entrar en la mente del hablante (o del escritor) es solo una parte del desafío. Eso nos devuelve a la pragmática y a nuestro amigo Goostman.

LAS MÁXIMAS DE GRICE PARA UNA BUENA CONVERSACIÓN

Podrías afirmar que los diseñadores de Eugene Goostman usaron la pragmática como una herramienta, como un arma. El código del bot conversacional explota una zona de la pragmática que pone del revés la comprensión del lenguaje a través del abuso del sarcasmo, el encubrimiento constante y copiosas distracciones. En otras palabras, explota la comprensión, asumida por los jueces, de que los muchachos ucranianos de trece años de edad a veces evitan las preguntas no solo porque desconozcan su respuesta, sino porque no les da la gana hacer el esfuerzo de contestarlas.

Los lingüistas reconocerán que Goostman viola alegremente las llamadas «máximas de Grice».[5] A principios del siglo XX, el filósofo del lenguaje Paul Grice ofreció cuatro máximas para mantener una conversación con éxito:

1. La máxima de cantidad. Intenta ser lo más informativo posible, ofrecer tanta información como sea necesario, pero no más.

2. La máxima de calidad. Intenta ser sincero, y no ofrecer información que sea falsa o que no pueda demostrarse.

3. La máxima de relevancia. Intenta ser relevante, y decir cosas que resulten pertinentes para el diálogo.

4. La máxima de modo. Intenta ser tan claro, tan breve y tan ordenado como puedas, y evita la oscuridad y la ambigüedad.

Eugene Goostman es un criminal reincidente. Viola todas las máximas de Grice

para sacarse de encima a los desafortunados jueces humanos que presumiblemente intentaban confirmar que tuviera una comprensión normal y la capacidad básica y la habilidad de conversar. Desde la buena voluntad, la gente suele adaptarse a los déficits que brotan del uso de un segundo lenguaje o de la desfachatez juvenil siempre y cuando haya un esfuerzo evidente por ser informativo, sincero, relevante y razonablemente claro. No sucede así con Goostman, que viola las máximas a propósito.

Pero, en una conversación cotidiana, que se violen las máximas de Grice indica a los interlocutores normales que pasa algo raro. Si alguien se te acerca en un café y te dice: «Me llamo Brian Johnson, ¿me puedes decir la hora?», la petición, aunque comprensible, te resultará extraña. Las máximas de Grice explican por qué: la pregunta viola las máximas de cantidad y relevancia, e induce a confusión al ofrecer una información que no es necesaria para averiguar la hora. (Nos preguntamos: «¿Por qué me ha dicho su nombre? ¿Acaso cree que tiene que mostrarme unas credenciales para preguntar la hora?».)

Volvamos a Goostman: en una prueba estructurada, de tiempo limitado, con el mundo mirando, es sencillamente ridículo violar las máximas de Grice de manera sistemática, cuando el objetivo mismo del test es determinar la inteligencia a través de una conversación normal y corriente. Y sin embargo funcionó, en el concurso al menos. No debería haber sido así. Y la reacción para que nos libremos del test de Turing también fue innecesaria y ridícula. Para librarte de Goostman tienes donde elegir: puedes usar tests simplificados de opción múltiple como los esquemas de Winograd, o mejorar las pruebas conversacionales (o basadas en monólogos) aplicando una pena severa (como la descalificación) a quien use técnicas de evasión. Por decirlo de otro modo, es fácil acabar con los trucos. El problema complicado de verdad es alcanzar la comprensión.

CUANDO LO RAZONABLEMENTE BUENO NO ES LO BASTANTE BUENO

El pobre desempeño obtenido en las pruebas con esquemas de Winograd deja al descubierto otra de las debilidades capitales de los enfoques contemporáneos,

basados en datos, de la IA. En un apartado de su artículo que llevaba el título de «El encanto de la estadística», Levesque dijo que el enfoque de la «simulación de datos» era como hacer algo «vagamente transformable en X» en oposición a hacer directamente X, siendo esa X «una de las muchas instancias del comportamiento inteligente». Esta, insistió, es «la cuestión dominante para la ciencia de la IA».[6] Pero Levesque podría haber ido un poco más allá con su crítica. Hacer algo «vagamente trasformable en X» —en otras palabras, simular inteligencia de cara a una tarea X— conduce de manera ineludible a una serie de problemas que requieren de una comprensión real del reino de lo irresoluble, sin que se conquisten esos escasos puntos porcentuales (o, por decirlo de otro modo, sin ascender por la pirámide lingüística del significado). Así se crea un escondrijo para unos problemas que necesitan una comprensión real. Y también se crea una confusión general.

El público, testigo de lo que en apariencia es una actuación impresionante, llega a la conclusión comprensible de que las brechas que quedan entre máquinas y seres humanos irán desapareciendo a medida que las técnicas informáticas y de análisis, cada vez más potentes (y cada vez con más datos), vayan ayudando a los sistemas del futuro a alcanzar un mejor rendimiento. Pero esa confianza ciega en que los problemas sin solucionar vayan a tener solución algún día ignora la distribución real de los problemas y sus soluciones, donde los de tipo sencillo aparecen con frecuencia en las simulaciones basadas en datos y los demás, de mayor rareza y dificultad, caen fuera del alcance de los análisis basados en frecuencias. Eso quiere decir que los problemas difíciles —las respuestas erróneas— quizá requieran de un enfoque nuevo por completo, y no de más datos. Ese enfoque nuevo y necesario, tal y como he comentado, presupone un armazón no deductivo o inductivo y que ponga el énfasis en la inferencia. De hecho, si pudiera realizar inferencias abductivas de sentido común con fiabilidad, DeepRead efectuaría también conjeturas correctas acerca del conocimiento y los hechos que se asumen y que no están presentes en la historia anterior. No puede deducirlos (no son problemas lógicos) y, puesto que no se pueden inferir a partir de la sintaxis (o los datos) de la historia, tampoco puede inducirlos. DeepRead espera que haya avances en la inferencia; por desgracia, las inferencias necesarias no se pueden programar en este momento. No hay nada hacia lo que crecer en escala —la inferencia que se necesita es distinta, y representa un descubrimiento conceptual futuro.

Me he centrado en los esquemas de Winograd porque claramente requieren de inferencias no inductivas a la vez que plantean lo que a juicio de la mente

humana parece una prueba de comprensión de una sola oración. Cuesta debatir la injusticia de una prueba que consiste en una simple pregunta. Los problemas de los esquemas de Winograd tratan de manera deliberada con objetos comunes, lo cual implica que cualquier nombre o sintagma nominal aparecerá por lo general en textos como pueden ser las páginas web (por ejemplo, «cocodrilo» o «trofeo»). No es necesario ningún conocimiento técnico. Pero la coincidencia de los nombres o sintagmas que aparecen en esa única pregunta hace que la frecuencia esperada caiga de manera notable, a veces hasta el cero, como en el caso de los cocodrilos y los cien metros vallas. Así, aunque directos, todos los ejemplos son relativamente raros en el big data. Y, en los casos en los que los dos nombres o sintagmas nominales puedan aparecer juntos en una frase de la red (como en la pregunta del trofeo y el maletín), con solo cambiar la relación entre ambos se recuperará esa infrecuencia —y se derrotará, tal y como hemos visto, a las tácticas contemporáneas centradas en datos—. Por tanto, a todos los efectos prácticos, las preguntas de los esquemas de Winograd no se pueden simular, y eso explica el pobre rendimiento de los sistemas que intentan automatizar la prueba.[7]

Del armazón inferencial que he descrito antes se deduce que los sistemas deductivo e inductivo no resultan adecuados para la inteligencia artificial general. También he explicado que ningún tipo dado de inferencia puede reducirse a otro tipo (recuerda los ejemplos de los silogismos), de modo que, por ejemplo, no podemos decir que la abducción sea una especie de inducción con extensiones adecuadas, ya que eso la convertiría en una forma simbólica (o «forma lógica») diferente por completo. Sigue en pie la pregunta sobre si alguna combinación de inducción y deducción podría ampliarse camino de la inferencia abductiva. Pero no es así, por el mismo motivo por el que no se puede subsumir un tipo dentro de otro: son distintos e implican capacidades en esencia diferentes. (A modo de comparación, podemos pensar lo siguiente: si sé hablar alemán y español, ¿podría de algún modo unir ambos idiomas y entender el ruso?) Sin embargo, a veces, los diferentes tipos de inferencia se combinan en el trabajo de la IA, y los resultados han sido interesantes.

Los equipos de investigación enfrentados a problemas tan difíciles como el procesamiento del lenguaje suelen construir amplios sistemas híbridos utilizando arquitecturas, bases de datos y algoritmos de predicción o inferencia especialmente diseñados para ello. De manera invariable, esos sistemas incluyen el aprendizaje automático como componente o subsistema. También utilizan unas bases de conocimiento y unas técnicas de inferencia que se remontan a la

época premoderna de la IA clásica. Los sistemas híbridos pueden obtener resultados impresionantes. Al no tener preferencias, se sirven de todos los métodos, técnicas y algoritmos disponibles para resolver problemas difíciles. Uno de esos sistemas en concreto fascinó a los medios y al público en el momento en que la IA misma emergía como tema candente en un mundo de cambio acelerado. Y es que, a primera vista en todo caso, el sistema pareció comprender —comprender en serio— cómo responder preguntas planteadas en inglés.

WATSON, QUERIDO WATSON

IBM tiene una habilidad especial para diseñar sistemas de juegos de IA que convierten a este legendario gigante tecnológico en la comidilla tecnológica. En 1997, IBM saltó a los titulares con Deep Blue, el superordenador que jugaba al ajedrez y que derrotó al por entonces campeón Garry Kasparov en un esperado acontecimiento televisado. Deep Blue causó sensación, en gran parte porque se suponía que ese punto de desvío en el que las máquinas vencerían a los seres humanos se encontraba a años de distancia. Los eruditos especularon con la posibilidad de que la inteligencia verdadera hubiera llegado a los superordenadores. Se volvió a hablar del test de Turing. Los comentaristas de la prensa se preguntaron, ansiosos, cuánto faltaría para que las máquinas nos superaran en todo lo demás.

En retrospectiva, el espectáculo de Deep Blue tuvo poco que decir acerca de la inteligencia de la máquina, pero anticipó el aluvión de sistemas basados en datos que iban a pasarse las décadas siguientes derrotando a los seres humanos en otros juegos. Al final, el ajedrez es un juego que se disputa según unas reglas deterministas. Garry Kasparov es un genio, sin duda, pero Deep Blue le derrotó a base de pura potencia informática: evaluando más movimientos, viendo a mayor profundidad las ramificaciones del juego. Los entusiastas de la IA llevan décadas quejándose de la tendencia del público, tras el bullicio inicial, a desdeñar con rapidez los nuevos éxitos por «no ser inteligentes de verdad», y Deep Blue no fue ninguna excepción. Con el cambio de siglo ya estaba en gran parte olvidado: otro hito que fracasó a la hora de dejar pasmado a un público que quizá percibió en lo más hondo que, después de todo, saber jugar al ajedrez no equivalía a

disponer de una inteligencia general. Deep Blue fue un escaparate para IBM, sus ingenieros bien financiados y sus veloces ordenadores.

En 2001, con el estallido de la burbuja cibernética, la excitación en torno a la IA se desvaneció durante un tiempo. Los miles de millones de dólares invertidos en programas y visiones se habían evaporado porque estos no lograron percibir una serie de realidades básicas. La IA parecía futurista e insegura en términos financieros, y eso era exactamente lo que inversores y emprendedores no deseaban, pues seguían lamiéndose las heridas. Aquel invierno duró poco. Para 2004, Google ya había madurado y las redes sociales estaban de camino. (Algún intento anterior, como el de Friendster, había fracasado, pero aquel concepto tan excitante se encontraba en el aire.) La web 2.0 estaba al caer. Aquel mismo 2004, la dirección de IBM comenzó a buscar un impulso en sus relaciones públicas. El ajedrez era agua pasada, pero la red estaba cobrando forma y revitalizando a todo el mundo con sus excitantes posibilidades. Los juegos seguían llamando la atención, y contaban con el beneficio añadido para las compañías tecnológicas de que la IA a menudo tenía éxito en las restricciones artificiales de un juego, por mucho que se fuera a pique ante el más simple sentido común. Y quiso la suerte que un juego popular en concreto estuviera ya en las noticias: Jeopardy!

El concurso de preguntas televisivo Jeopardy! es una especie de versión lúdica del test de Turing simplificado (o eso parece). Se trata de una prueba amplia de conocimiento de hechos en la que se invierte la «conversación», de modo que los concursantes reciben una respuesta y deben ofrecer la pregunta correcta. Por ejemplo, para la propuesta «Desarrollado por IBM, batió a Kasparov al ajedrez», la respuesta correcta sería «¿Qué es Deep Blue?». En 2004, los índices de audiencia de Jeopardy! estaban por las nubes gracias a una persona, Ken Jennings, el campeón que regresaba al programa después de batir el récord en él con sus 74 victorias consecutivas. IBM desafió a su equipo de investigación para que desarrollara un sistema que pudiera batir a Jennings, o a cualquier otro campeón. Aquello fue un sueño para los investigadores, que en efecto recibieron carta blanca para extender los límites de la IA.

La dirección de IBM vio la oportunidad de obtener un beneficio económico, por supuesto, y una oportunidad buenísima (o una tragedia) en términos de relaciones públicas. Pero al parecer a la dirección también le había picado el bicho de la ciencia ficción, la idea futurista de un sistema de comprensión del lenguaje «privilegio de IBM» que hiciera sonar la bocina antes que los seres

humanos, recitara del tirón todas las respuestas y dejara a los espectadores boquiabiertos en sus casas. Volvía a ser lo de Deep Blue y el ajedrez, pero a la vez no lo era. Jeopardy! se basa en el lenguaje. Es un sistema de pregunta-respuesta, en sentido técnico, con la simplificación de que la pregunta se presenta en una frase sencilla y la respuesta siempre comienza por «¿Qué es...?» y acaba con la solución en una frase simple. Eso parece bastante más ligado a la inteligencia humana, y nos recuerda al test de Turing. Jeopardy! pareció una oportunidad para dar un gran salto en la carrera armamentística de la IA. (Hasta una mosca que pasara por allí se habría excitado aquel día de 2004, en el Centro de Investigación de IBM, al oír a los directivos de la compañía: «Vamos a hacer historia. El dinero no es el objetivo. Encontrar la manera de construir un sistema que juegue al Jeopardy!».)

IBM reunió un equipo de gran talento y se pasó los siguientes tres años realizando las diligencias debidas sobre el requisito de un sistema de pregunta-respuesta que fuera capaz de jugar al más alto nivel del Jeopardy! Eso implicaba que las preguntas fueran de dominio abierto, porque el juego cubre temas diversos. (Es algo que forma parte de su atractivo.) Al parecer, no se podía piratear ni codificar el sistema para que se centrara en temas específicos. Bautizaron a aquel sistema futuro como Watson (en honor del fundador y primer director general de IBM, no del compinche ficticio de Sherlock Holmes).

Un sistema más antiguo llamado PIQUANT les proporcionó una pequeña ventaja. IBM lo había construido en 1999 para que compitiera en los congresos de recuperación de textos (TREC, en sus siglas inglesas) bajo patrocinio del Instituto Nacional de Estándares y Tecnología (NIST, en inglés). PIQUANT había rendido muy bien y de manera consistente en los concursos de los TREC, pero estos consistían en juegos simplificados de pregunta-respuesta, y Jeopardy! representaba, por contraste, un desafío más amplio y vasto. Por ejemplo, PIQUANT respondía a las preguntas con un conjunto predeterminado de etiquetas, del tipo PERSONA, LUGAR, FECHA o NÚMERO. Dado el texto que se le suministraba al sistema como pregunta, PIQUANT generaba la etiqueta que representaba el tema pertinente. El juego de Jeopardy! abarca un amplio número de temas.

No fue posible ampliar PIQUANT para aquel desafío, y se presentaron otros problemas. En Jeopardy!, el concursante no debe apretar el botón a menos que esté muy seguro de acertar. Los errores se penalizan. Así que el tipo de pregunta-respuesta que se necesitaba para el Jeopardy! informático tenía diferencias

importantes. Eso significó, entre otras cosas, que el sistema iba a necesitar un rediseño completo. El nuevo sistema tenía que «hacerse» al juego del Jeopardy! a través del prueba-error constante, mientras que PIQUANT era un sistema de pregunta-respuesta universal.

Dave Ferrucci, responsable del proyecto Watson en IBM, se encargó de aquel rediseño completo y no tardó en reconocer que el Jeopardy! —que seguía siendo un juego, no lo olvidemos, y no una lectura abierta— necesitaría un sistema que no se atormentara demasiado con el contexto en el lenguaje natural: «Ten en cuenta la expresión "¡Qué obra tan mala!". ¿A qué se refiere con "obra"? ¿Una obra de teatro? ¿Un edificio en construcción? Y, en esta frase, ¿"mala" significa buena? Es evidente que hacía falta un mayor contexto para interpretar el significado pretendido de manera acertada».[8]

Ferrucci también conocía las ideas del pionero de la IA Marvin Minsky, según las cuales los problemas complejos e inteligentes se han de resolver de manera políglota, por parte de cerebros (u ordenadores) con diversos submódulos que descompongan esos problemas difíciles en fragmentos más manejables. La tarea, a continuación, consiste en combinar las diferentes respuestas de los módulos en una respuesta o solución «global». Minsky denominó ese método hacia la IA como «la sociedad de la mente», y Ferrucci lo adoptó como inspiración para diseñar a Watson.[9]

El equipo de Watson desarrolló DeepQA, un sistema de búsqueda de respuestas que generaba varias respuestas posibles a las preguntas y ofrecía la mejor basándose en un análisis múltiple. En esencia, DeepQA implementó una versión de software de la idea de la sociedad de la mente para el juego del Jeopardy! El sistema minimizaba las respuestas demasiado rápidas segmentando la pregunta y postergando la respuesta ganadora hasta haber reunido todas las pruebas a su disposición. Una pieza del rompecabezas para responder una pregunta podría estar en el tramo final de este análisis de DeepQA, por ejemplo. La arquitectura de Watson formó parte de su «inteligencia» final y fue uno de los motivos por los que, tras años de desarrollo, Watson acabó convirtiéndose en una importante solución híbrida para la complejidad del problema basado en el lenguaje.

Hubo otras razones. Para comenzar, se dedicó una enorme cantidad de análisis humanos al diseño, desarrollo y comprobación de Watson. Esa contribución humana a los diseños de sistema exitosos suele pasarse por alto, en especial cuando se supone que la aplicación ha de ser un escaparate para la IA. De hecho,

está claro que Watson fue resultado tanto de un análisis del juego cuidadoso y perspicaz por parte del equipo de ingenieros como de las mejoras que se realizaron sobre PIQUANT y los sistemas de búsqueda de respuestas en general. Se analizaron miles de partidas de Jeopardy! Primero se estableció un rendimiento de referencia modificando PIQUANT para que jugara al Jeopardy! (Como era predecible, le fue muy mal. El sistema PIQUANT obtuvo un deprimente 16 % de respuestas correctas sobre el 70 % de preguntas a las que contestó, a lo que el equipo de IBM pasó a referirse como «16@70».) A continuación, rediseñaron el módulo de DeepQA, centrándose en las preguntas de Jeopardy! para encontrar pistas específicas que indicaran posibles respuestas. Ese proceso de prueba y error para adecuar la segmentación de DeepQA específicamente a Jeopardy! resultó capital para su éxito definitivo. Los ingenieros de IBM incluso le pusieron un nombre: AdaptWatson.

AdaptWatson no era Watson; era el proceso humano para mejorar a Watson apuntando hacia los trucos específicos de Jeopardy! e incluyéndolos en DeepQA. En total se diseñaron, implementaron y refinaron más de un centenar de módulos de procesamiento del lenguaje con propósitos especiales usando el protocolo de AdaptWatson. Se trató de un inmenso proceso de trabajo que en su cenit involucró a 25 investigadores, incluyendo a ayudantes procedentes de las universidades de la zona. Desde el punto de vista de su ingeniería, AdaptWatson fue notable: se jubiló a PIQUANT y una segmentación «óptima para Jeopardy!» fue tomando forma en torno al desempeño en el juego. DeepQA se convirtió en un efectivo circuito de retroalimentación que implicó a docenas de expertos humanos, y estos se dedicaron a afinar el sistema Watson para que pudiera jugar al más alto nivel de Jeopardy! (Quizá se valorara menos el hecho de que el sistema resultante fuera más débil —consecuencia directa de esa estrategia de ingeniería orientada de manera específica a los juegos.)

Inspirado en la idea de la «sociedad de la mente», DeepQA incluyó innovaciones en el procesamiento de lenguajes naturales (PLN, o NLP en inglés). Los sistemas de búsqueda de respuestas más antiguos, como PIQUANT, dependían de procesos de segmentación parecidos al que utilizó el Watson de IBM: primero analizaban la pregunta, a continuación buscaban y recuperaban las posibles respuestas, seguidamente les daban una puntuación y acababan ofreciendo la mejor. Sin embargo, PIQUANT y otros sistemas de la generación anterior tienen una falla en común: se pueden quedar atrapados en las respuestas incorrectas por culpa de errores cometidos en una fase temprana del análisis. El equipo de Watson modificó una vieja arquitectura desarrollada por IBM entre 2001 y 2006,

y conocida como Aplicaciones para la Administración de Información No Estructurada (UIMA, en sus siglas inglesas). Las UIMA son de instalación automática: los módulos de software se pueden someter a la segmentación, retirarse, modificarse, y someterse de nuevo. De igual manera se pueden intercambiar algoritmos enteros, de modo que la segmentación se acelere y facilite las ejecuciones de prueba y error, y las pruebas exhaustivas que se necesitan en proyectos tan amplios y complejos. Al usar las UIMA, el equipo de Watson produjo una segmentación más sofisticada para jugar a Jeopardy!, que AdaptWatson no dejaba de afinar. El sistema híbrido funcionó: Watson comenzó a jugar al más alto nivel de Jeopardy!

No obstante, al igual que Deep Blue, su antecesor de IA débil, Watson se diseñó desde un primer momento para que se le diera bien una sola cosa. Resulta comprensible que los detalles del sistema Watson sean complejos, pero incluso una breve ojeada al desarrollo del sistema y a sus componentes clave disipa cualquier pretensión de alcanzar una comprensión real. Por ejemplo, Deep QA depende de un cajón de sastre de técnicas relativamente bien comprendidas de la investigación en PLN, como la disección de frases, la resolución de algunos tipos de correferencia (resolver el antecedente del pronombre, como hemos visto antes), el reconocimiento de entidades nombradas como personas o lugares, y la clasificación de las preguntas mismas por tipos, incluida la categoría de «trivialidades». El equipo de Watson usó esas técnicas pero las aplicó de manera específica al juego de Jeopardy! También redactaron un módulo de «secciones de pregunta», que busca restricciones evidentes y específicas sobre las respuestas en las preguntas mismas. Por ejemplo, la frase «Esta palabra de cuatro letras significa...» en una pregunta señala que su respuesta será una palabra de cuatro letras. Una vez más, estamos ante un diseño y un desarrollo destinados a optimizar el desempeño en un juego, no para que la máquina alcance la comprensión de un lenguaje natural general.

Echar un vistazo más meticuloso a los resultados de la búsqueda de respuestas del sistema Watson quizá haga que su cualidad en apariencia mágica se desvanezca con mayor rapidez. El seguimiento y la recuperación de posibles respuestas es un éxito en gran medida gracias a la explotación de un atajo en Jeopardy! que fue descubierto por los ingenieros que usaban AdaptWatson para identificar problemas y posibles mejoras: el 95 % de todas las respuestas a las preguntas de Jeopardy! son títulos de Wikipedia. Ese descubrimiento afortunado posibilitó todo aquel esfuerzo; con solo emparejar las preguntas a los títulos de Wikipedia se obtendría un desempeño sobrehumano. Para variar, el diablo se

esconde en los detalles.

El sistema Watson es impresionante pero, tal y como sucedió con el anterior éxito de IBM, el programa Deep Blue para jugar al ajedrez, no queda claro si hemos de dirigir nuestras loas a sus recursos superinformáticos (Watson usó más de doscientos servidores de ocho núcleos) o a la perspicacia y diligencia de sus ingenieros humanos —quienes, al fin y al cabo, contaron con una buena financiación para trabajar de forma especial en ese beneficio excepcional para el marketing de IBM—. Hay que reconocerle al equipo que Watson se diseñó para realizar otras búsquedas abiertas en fuentes desestructuradas como blogs, biblias digitales y demás. Y Wikipedia misma está en gran parte desestructurada, aunque existe una base de datos llamada DBPedia, y Watson la incluyó. Pero, de nuevo, se trató de una búsqueda específica de representaciones informáticas de preguntas, y devolvió pasajes que debían ser analizados con técnicas que llevan décadas utilizándose en la investigación sobre el procesamiento del lenguaje: rellenar los huecos en la pregunta con palabras y frases supuso una buena opción, a juzgar por los resultados de la búsqueda. No es poca cosa, desde luego, pero su valor se ve reducido, una vez más, por la constatación de que el análisis de textos más abierto de Watson fue también el menos fiable —otra observación capital y desmitificadora—. Cuando no había un emparejamiento con los títulos de Wikipedia, las aptitudes del sistema Watson resultaron mucho menos impactantes (pero claro, se trataba solo del 5 % de las respuestas). Resulta que el propio Jeopardy! tiene su truco, porque es un juego de trivialidades y las trivialidades son recuperables. Este rasgo del juego, descubierto a través del protocolo humano AdaptWatson, explica mejor que ninguna otra cosa el desempeño sobrehumano que este acabó teniendo —también, el motivo por el que la incursión de IBM en el mundo de la salud ha sido sin duda menos exitosa.

De manera predecible, el éxito de Watson dio pie a una discusión acerca de la madurez de una IA que entendiera de verdad el lenguaje natural. Sin embargo, Ferrucci y su equipo descartaron de buenas a primeras una tarea mucho más simple que jugar a Jeopardy!: la de leer el periódico (desde luego, más simple para las personas). Ferrucci lo dijo de manera explícita, señalando que incluso el contenido de las páginas de Wikipedia (no solo sus títulos) generaba problemas a la IA de tal complejidad que la conversión con propósitos generales de textos abiertos en un formato que pudiera leerse informáticamente era irresoluble —se consideró aquel objetivo de manera breve y acto seguido se abandonó—.[10] En cambio, el equipo de Watson identificó un conjunto de objetivos de gran valor en las fuentes de información —extractos y fragmentos que contenían respuestas

probables— que encajaban en el proceso de segmentación de Watson y que elevaban el número de posibilidades de obtener las respuestas correctas. Al optar por ese enfoque, el equipo jubiló a todos los efectos el juego de Jeopardy!, tal y como había hecho antes con el ajedrez. Pero una vez más se demostró esa máxima cada vez más evidente: que toda IA exitosa es una IA débil. (Y demostró también el corolario: el éxito en un juego provoca la excitación del público, pero no representa ningún avance hacia la inteligencia artificial general.) Watson no es un paso en la dirección de la inteligencia general, sino una prueba más de que la búsqueda de la generalidad sigue enredada en el misterio y la confusión. Mientras que el equipo de IBM obtuvo una victoria impresionante sirviéndose de un potente sistema híbrido, no descubrió la clave de la comprensión del lenguaje. Todos los problemas de la programación de inferencias inductivas para llegar a la inteligencia general continúan estando ahí.

LA TRAMPA DE LA DEBILIDAD Y EL LENGUAJE

Me he centrado con cierto detalle en la comprensión del lenguaje (en vez de, pongamos, los problemas de la robótica) porque revela de manera muy evidente la trampa de la debilidad en la que caen todos los sistemas conocidos. A partir de ahora voy a reseñar algunas aplicaciones recientes que publicitan una capacidad para entender el lenguaje de la que carecen. Más bien, son ejemplos de IA débil que con demasiada asiduidad se disfrazan de algo más.

Por ejemplo, el programa Duplex de Google. Lanzado en 2018, Duplex realiza llamadas en nombre de su dueño para cumplir con tareas rutinarias como hacer reservas y concertar citas. Duplex cuenta con una voz que parece humana (lo que más adelante, cuando aumentó la presión del público, llevó a Google a advertir del carácter automatizado del sistema ante quienes llamaban). Duplex parecía el advenimiento de HAL, de la película 2001: Una odisea del espacio, hasta que nos enteramos de que el sistema, desarrollado con los amplísimos recursos informáticos y de datos que tiene Google, prometía solo hacer reservas en restaurantes, pedir citas en peluquerías y averiguar las horas de apertura de unos pocos negocios selectos. Parece poco, y la cosa empeora. Tras la fase de demostración, Duplex llegó a los móviles Android sin la opción de hacer reservas abiertas ni de preguntar sobre horarios de apertura. Solo hacía reservas

en restaurantes. En palabras de Marcus y Davis, «cuesta encontrar algo más restringido».[11]

Junto con Duplex ha habido multitud de propuestas recientes que, sirviéndose del big data y el aprendizaje automático, han prometido aptitudes propias de HAL para acabar cayendo también en la trampa débil de la IA. Los asistentes virtuales de voz como Siri, Cortana, Google Assistant y Alexa responden a las preguntas que les plantea la gente y pueden incluso darse a la broma, como respuestas humorísticas a preguntas juguetonas (o insultantes). Pero su comprensión del lenguaje natural es una fachada, como sabrá cualquier persona que haya interactuado con ellos. Igual que Watson, lo que se les da mejor son las trivialidades que pueden seleccionar en las bases de datos y paneles informativos (como Wikipedia) de la web. «¿Quién ganó la Super Bowl de 1975?» es una buena pregunta, pero «¿Puede un tiburón jugar a las damas?», no (recordemos los esquemas de Winograd).[12] En general, las preguntas que salen a explorar, aunque sea de manera ligera, bajo la capa de hechos extraídos de la red, algo para lo que se necesita un conocimiento y una comprensión reales confunden una y otra vez a esos sistemas. Su capacidad comienza y acaba en esa capa fina como el papel de los hechos recopilados y risas enlatadas. Igual que Goostman, carecen de una comprensión real, así que no pueden conectar de verdad con nosotros —y, a menudo, tampoco pueden ayudarnos.

En los sistemas como Watson, que abordan el lenguaje natural, lo limitativo es una característica endémica. Como hemos visto en capítulos anteriores, esto se debe a que acciones como la lectura y la conversación son en realidad hitos de la inferencia profundos y abiertos, que requieren de la comprensión del mundo que nos rodea. El Talk to Books de Google, que Ray Kurzweil presentó a bombo y platillo, como era de prever, durante una charla TED de 2018, prometía una capacidad de búsqueda de respuestas sin igual a través, tal y como dijo Quartz, de la «lectura de miles de libros».[13] En realidad, el sistema catalogó cerca de cien mil libros, codificando sus frases de manera numérica a través de vectores (estructuras de datos) y usando el aprendizaje automático (como si hubiera otra cosa) para calcular su similitud con otros vectores. Por enésima vez, se trata de una versión elegante del supuesto de frecuencia y de la restricción empírica.

Tras esa ostentosa demostración en clave de TED, las numerosas limitaciones de Talk to Books no tardaron en salir a la superficie. En su base de datos se podían recuperar detalles y anécdotas, pero rara vez contestó a alguna pregunta que requiriera de una inferencia real, como la abducción. En caso de catalogar El

gran Gatsby, por ejemplo, el sistema podía responder a una consulta sobre su autor (F. Scott Fitzgerald) o incluso sobre el nombre de pila de Gatsby (Jay). Pero las preguntas sencillas sobre el argumento o los personajes, que necesitan de inferencias basadas en el conocimiento, superaban con rapidez la capacidad del sistema. Habiendo leído la novela, es sencillo contestar a la pregunta «¿En qué ciudad conoce Gatsby al protagonista?» Pero, para contestarla, Talk to Books habría necesitado un poder inferencial más allá de su alcance. Marcus y Davis le preguntaron «¿Dónde se conocieron Harry Potter y Hermione Granger?» y obtuvieron respuestas que ni siquiera aparecían en Harry Potter y la piedra filosofal, y que (peor incluso) omitían el tema central de la pregunta: una localización; en concreto, aquella en la que se conocieron los personajes.[14]

El desempeño restrictivo es un problema de todos los sistemas que abordan la comprensión del lenguaje, y no solo del Talk to Books de Google. El lenguaje trata el mundo que hay «ahí fuera», y eso implica la necesidad de un conocimiento y de una cierta comprensión de lo que significan las cosas. Como hemos visto también, la trampa de lo restrictivo es consecuencia del propio enfoque basado en datos, desde su concepción misma. Parafraseando al personaje cómico y cinematográfico Ace Ventura, «Débil es la inteligencia artificial. La inteligencia artificial es débil».

Hay otros ejemplos ya célebres —o quizá deberíamos decir tristemente célebres —. En 2016, Microsoft lanzó su esperadísimo bot conversacional, Tay. El gigante del software publicitó a Tay como un salto espectacular sobre los sistemas antiguos basados en reglas, como aquella famosa ELIZA de apariencia humana salida de los años sesenta, ya que de hecho podría aprender de la interacción con el usuario y los datos online. Pero diríase que no se estudiaron bien las lecciones de la inducción y sus límites, pues Tay se tragó feliz la secuencia de clics racistas y sexistas con que lo trolearon, además de otros discursos de odio que encontró en la red. Tay se convirtió en un alumno aventajado en metadatos, dedicándose a soltar tuits que decían «Es que odio a las feministas, joder» o «Hitler tenía razón: odio a los judíos» para consternación de Microsoft, que tardó menos de un día en cancelar esa exhibición de odio.[15] Pero deberían haber previsto ese resultado, dada la naturaleza esencial que se escogió para su diseño, basada en el concepto de «basura entra, basura sale» (GIGO en sus siglas inglesas). Tay fue un ejemplo de miopía corporativa acerca del propio enfoque técnico —y un ejemplo más de IA débil—. En este caso, una comprensión real habría otorgado a Tay un mínimo de capacidad para filtrar aquellos tuits que resultaran ofensivos. Pero, puesto que para comenzar no

disponía de esa comprensión real sobre el lenguaje o los tuits, se puso a regurgitar todo lo que consumía. Tay es un ejemplo memorable (pero, por desgracia, fácil de olvidar) del carácter de sabio idiota que tiene la IA basada en datos.

La comprensión del lenguaje natural puede resultar difícil, pero al parecer también es irresistible. Facebook se puso a la cola de las decepciones al caer con el anuncio de un sistema que podría leer «una sinopsis de El Señor de los Anillos y responder preguntas al respecto», en palabras de Technology Review. Pero esa sinopsis consistió en cuatro líneas de oraciones simples como «Bilbo viajó a la cueva. Gollum había dejado caer el anillo en ella. Bilbo cogió el anillo». Y el sistema solo pudo contestar a las preguntas que estuvieran presentes en las oraciones, como «¿Dónde está el anillo?», etc. Las preguntas que reclamaran una comprensión del texto no eran posibles. En general, no se pudo responder a nada que se preguntara un porqué —pongamos, «¿Por qué viajó Bilbo hasta la cueva?»—. El sistema reduce El Señor de los Anillos a unas pocas líneas de texto, a una sinopsis, y contesta solo a las preguntas más prosaicas y estúpidas, sin demostrar la menor comprensión por el camino. La restricción es integral.

El armazón de la inferencia ayuda a dar un sentido a la trampa de la restricción. Watson, un sistema que hemos examinado con cierto detalle, acabó ofreciendo una actuación impactante en un juego complejo que entrañaba un tratamiento inteligente de información desestructurada —principalmente, de páginas web, y en especial, de páginas de Wikipedia (y, en ella, sobre todo de sus títulos)—. Una inmersión más profunda en el sistema nos revela su diseño híbrido, con reglas codificadas y métodos estadísticos para obtener respuestas, pasando por el árbol de búsqueda Monte Carlo de cara a la apuesta gemela del día y el Jeopardy! final (en el comentario anterior no tratamos el tema de las apuestas). En términos de inferencia, el análisis de las preguntas de Watson confió sobre todo en las técnicas tradicionales del etiquetado por categoría gramatical, entidad y demás información —en otras palabras, reglas o técnicas de inspiración deductiva—. Las secciones de pregunta, como «Esta palabra de cuatro letras significa...», típicas de Jeopardy!, son de fácil manejo sin necesidad de estadísticas. Y se podrían realizar comentarios similares a otros aspectos del análisis de preguntas que se pueden identificar de manera fiable a través del examen de las preguntas. ¿Por qué no usar el aprendizaje automático? Pues porque muchos problemas que resultarían sencillos para un enfoque basado en reglas o de lógica deductiva plantean problemas irresolubles para el aprendizaje automático. Así que Watson fue un híbrido inteligente. Y, sin embargo, de

manera demostrable, quedó atrapado de todos modos en la trampa de la restricción. La explicación predominante de la trampa dice simplemente que la inferencia general, que no está disponible, tampoco se puede generar a través de la combinación de enfoques basados en reglas o en el aprendizaje. Por decirlo de otro modo, a falta de una inferencia abductiva, el desempeño del sistema ha de ser restrictivo —no disponemos de inteligencia general—. La restricción es inevitable.

Conste que Watson empleó un conjunto impresionante de técnicas de aprendizaje automático. Se analizaron unas veinticinco mil preguntas de Jeopardy!, que se convirtieron en 5.700.000 ejemplos de entrenamiento para el sistema. Este generó pares de pregunta-respuesta, acumuló pruebas en el proceso de segmentación y marcó la lista de pares usando técnicas estadísticas —todo lo cual fue posible gracias a los datos de partidas anteriores, de resultado conocido.

Resulta notable que Watson no se sirviera del aprendizaje profundo —al menos, no la versión de Watson que superó a los campeones humanos en el evento televisado de 2011—. El aprendizaje profundo no le habría ayudado —y aquí, de nuevo, hemos de reconocer el ingenio del equipo de IBM—. Se usó la regresión logística regularizada, una técnica relativamente simple de aprendizaje automático, pese a la disponibilidad de algoritmos de aprendizaje más potentes. (En 2011, el aprendizaje profundo seguía siendo bastante desconocido.) Unos sistemas de aprendizaje de mayor potencia no habrían hecho más que generar un mayor entrenamiento computacional y ensayos más costosos —al fin y al cabo, la AI es un juego de herramientas—. El sistema Watson no presentó ninguna innovación real respecto a ninguna técnica en particular, pero, al combinarlas en un marco que se pueda segmentar inspirado en la «sociedad de las mentes» (usando las UIMA), obtuvo unos resultados de primer orden. En términos inferenciales, Watson quizá sea el mejor ejemplo hasta la fecha del poder resultante de usar todos los enfoques deductivos e inductivos que la IA tiene a su alcance para combinarlos en una arquitectura inteligente. Pero quítale el truco de los títulos de Wikipedia y no habría tenido éxito. En última instancia, sigue siendo restrictivo —muy restrictivo—, igual que todos los sistemas conocidos, híbridos o no.

Podríamos acuñar otro término para explicar todo esto: llámalo la «trampa de la inferencia». Puesto que ninguno de los tres tipos conocidos de inferencia se puede reducir a alguno de los otros dos, pues son distintos, y se necesita la inferencia abductiva para llegar a la inteligencia general, las técnicas de

inspiración puramente inductiva como el aprendizaje automático siguen resultando inadecuadas, tanto da la velocidad que alcancen los ordenadores, y los sistemas híbridos como Watson también se quedan cortos en comprensión general. En los escenarios abiertos que necesitan un conocimiento del mundo, como la comprensión del lenguaje, la abducción es central e irremplazable. Por ello, los esfuerzos por combinar estrategias deductivas e inductivas estarán siempre condenados al fracaso —a lo sumo, como en el caso de Watson, podríamos tardar un poco más de tiempo en averiguar el motivo—. La disciplina necesita una teoría fundamental de la abducción. Mientras tanto, no dejaremos de caer en las trampas.

PIANOS LÓGICOS Y VIAJES

Charles Sanders Peirce conoció la posibilidad de usar máquinas para explorar la inferencia lógica. Los ordenadores no existían aún, pero sí había ideas al respecto, y se habían construido algunos artefactos. El filósofo y lógico británico John Venn (creador de los epónimos diagramas de Venn) especuló con la idea de construir una máquina lógica que fuera por completo automática. Y Allan Marquand, uno de los alumnos de Peirce en la universidad Johns Hopkins, comenzó de hecho a trabajar en una máquina lógica en 1881. En efecto, Marquand empezó a ampliar el protoordenador conocido como Piano Lógico de Jevons, bautizado en honor de su inventor, el inglés William Jevons. La máquina de Marquand debía resolver problemas de lógica deductiva, un apartado a cuyo estudio Peirce dedicó gran parte de su vida. Peirce mismo se interesó de manera activa en el desarrollo de la máquina lógica, y bosquejó diseños para las operaciones electromagnéticas del artilugio de Marquand.

Al escribir acerca de la experiencia en un artículo extrañamente clarividente, que tituló «Máquinas lógicas» y publicó en el American Journal of Psychology, Peirce comienza, como era característico en él, con una advertencia: «En el "Viaje a Laputa" aparece la descripción de una máquina que hace evolucionar la ciencia de manera automática —escribe—. La intención consiste en ridiculizar el Órganon de Aristóteles y el Órganon de Bacon demostrando la absurdidad de suponer que cualquier "instrumento" podría realizar el trabajo de la mente». Peirce, el escéptico, sin duda apreció la sabiduría de la imaginación de Swift.

Pero se dejó llevar lo suficiente por el espíritu prometeico como para subrayar la importancia del trabajo que Marquand y él habían emprendido. Reconoció los méritos de su pupilo y del predecesor de su pupilo: «Sin embargo, las máquinas lógicas de Jevon y Marquand son molinos que se alimentan de premisas y que arrojan conclusiones con el giro de una manivela». El inventor norteamericano Charles Henry Webb también había diseñado una máquina para que realizara operaciones aritméticas, y el genio inglés Charles Babbage desarrolló (junto con su protégée, Ada Lovelace) la prueba de concepto de una máquina más visionaria, que realizaría cálculos generales. Se trataba de máquinas que podrían «efectuar razonamientos en absoluto sencillos».[16]

A continuación, «Máquinas lógicas» se sume en un debate detallado sobre la automatización de los silogismos deductivos. Al final del artículo, para atar sus cabos sueltos, Peirce comenta la posibilidad de lo que ahora conocemos como inteligencia artificial: «Toda máquina dotada de razón, es decir, toda máquina, presenta dos motivos inherentes de impotencia. En primer lugar, está desprovista de toda originalidad, de toda iniciativa. No puede encontrar sus propios problemas, no puede alimentarse sola. No puede dirigirse entre los diferentes procedimientos posibles».[17] A continuación, Peirce cita un complejo problema lógico cuya solución requiere de la selección de unas premisas a través de docenas de pasos. Quizá ese ejemplo se pueda resolver de alguna manera, de algún modo (lo más probable es que se pueda resolver hoy en día). Peirce anticipa esa posibilidad, pero da igual. No es la cuestión. «Y, aunque lográramos hacerlo, seguiría siendo cierto que la máquina carecería por completo de iniciativas originales, y realizaría solo las cosas especiales para las que la habrían calculado.»[18] Como sucede con gran parte de su pensamiento, Peirce se sumó aquí a un debate que no se iniciaría de verdad hasta varias décadas después de su muerte.

Peirce añade también una idea simple que sigue rodeada de confusión, tiñendo la ciencia de mitología. «Esto, no obstante, no es ningún defecto en la máquina; no queremos que se dedique a sus asuntos, sino a los nuestros.» La trampa de la restricción también fue para Peirce un rasgo evidente de las máquinas: «La capacidad de la máquina tiene unas limitaciones absolutas; la han planeado para que haga algo en concreto, y nada más que eso».[19] Aquel científico, que dedicó la obra de una vida a explorar el misterio de la inteligencia humana, sabía demasiado bien que las máquinas eran, por diseño, unos sustitutos pobres e inadecuados. Las fantasías de Swift eran sabias.

En el siguiente siglo, Turing propuso que aceptáramos el desafío de imbuir las máquinas con una «iniciativa original», y eso debía comenzar con programarlas para que hablasen con nosotros. Turing era consciente de la objeción de Peirce, que en su artículo de 1950 atribuyó a lady Lovelace. También había jugado con algoritmos simples de aprendizaje, y en la década de 1950 apareció la red neuronal monocapa (llamada «perceptrón»). De manera comprensible, Turing quizá creyó que podría escapar a las objeciones de Peirce y Lovelace creando máquinas de aprendizaje según el modelo del cerebro humano. Leyendo «Maquinaria computacional e inteligencia», uno tiene la impresión de que el aprendizaje representaba la única huida real a las limitaciones inherentes a las máquinas, y la única esperanza real de pasar el test de Turing.

No lo ha sido, no ha sucedido. Pensar que pasará, que debe pasar, tiene unos efectos para la sociedad que se han vuelto demasiado evidentes. En la última parte de este libro vamos a observar algunas de las consecuencias del mito de la inevitabilidad, en especial, su efecto perjudicial para la ciencia misma.

Tercera parte

EL FUTURO DEL MITO

Capítulo 15

Mitos y héroes

Las ideas tienen consecuencias. En los próximos capítulos espero demostrar que las consecuencias del mito de la inteligencia artificial plantean una amenaza grave y trascendental para el futuro del descubrimiento y la innovación científicos —e, irónicamente, para el progreso en el ámbito mismo de la IA—. Esta parte final trata sobre nuestro futuro, pero debemos iniciarla en el pasado, ya que el problema de la creación de la vida, del diseño de una inteligencia artificial —en sentido literal, una mente en el interior de una máquina—, se ha visto imbuido siempre por un sentido mitológico, donde los humanos iban más allá de sus posibilidades y alcanzaban un poder divino. El mito de la IA es prometeico.

EL MITO DE PROMETEO

Prometeo robó a Zeus el fuego, que representa la vida, y lo usó para asar carne para todos los habitantes de la tierra. A nadie sorprenderá que Zeus se enojara y bajara a la tierra para exigir una parte justa de cada animal que la raza humana cocinara. Entonces, Prometeo engañó a Zeus para que escogiera como su porción solo las entrañas y las vísceras.

Zeus hizo lo que hacen siempre los dioses cuando alguien intenta usurpar su poder y su autoridad: castigó a Prometeo atándolo a un acantilado rocoso y mandando a un águila (el emblema de Zeus) para que se comiera su hígado. Cada noche, el hígado de Prometeo volvía a crecer; cada día, el águila regresaba para comérselo de nuevo.

Este relato trata la expansión de las facultades humanas y da testimonio del espíritu creativo de las personas, inherente y en apariencia inagotable. También

es una historia de soberbia. Prometeo podría haber conservado el fuego y el hígado ofreciéndole los mejores cortes a Zeus. Nuestro anhelo profundo por una IA verdadera halla inspiración en el mito prometeico. Queremos robarle el fuego a los dioses pese a que las consecuencias serán potencialmente horribles —el castigo eterno, ni más ni menos.

Prometeo fue un héroe castigado, y ese es el motivo por el que Mary Shelley tituló su novela imperecedera como Frankenstein o el moderno Prometeo. Con el paso de las décadas, Frankenstein se ha hollywoodizado y transformado en el relato bufo de un monstruo verde, pero en realidad es una historia sobre el espíritu prometeico en los seres humanos y sus consecuencias. Mary Shelley era muy joven y acababa de casarse cuando concibió a Frankenstein durante una pesadilla, en un hotel suizo, tras haberse quedado hasta tarde conversando con su marido, el poeta Percy Shelley, y con el también poeta Lord Byron. Sus imágenes oníricas no trazaron un plan para crear al monstruo, así que su creación planteó desde el principio la pregunta humana en el corazón del mito moderno: ¿y si un ser así fuera posible?

A través de una variedad no especificada de métodos ocultos, como la galvanización de los tejidos muertos, el doctor Frankenstein obra el milagro prometeico: crea vida inteligente a partir de la materia muerta. La historia es antecesora de retratos posteriores y específicamente mecánicos sobre la creación de una inteligencia usando la ciencia y la tecnología, y además es —y esto es importante— una historia de gran calado humano sobre el aislamiento espiritual. El doctor Frankenstein es un científico loco que posee conocimientos ocultos y prohibidos, lo cual le permite jugar a ser Dios. Su criatura cobra vida; tiene conciencia y anhela una pareja romántica. De manera inevitable, el mundo del doctor Frankenstein se desmorona, igual que el de Prometeo. A los diecinueve años, Mary Shelley captó aquel mito de la antigüedad y le devolvió la vida en el mundo moderno, igual que Percy Shelley cuando escribió posteriormente su famoso Prometeo liberado, una historia de redención. Los escritores románticos prestaron voz con efectividad a la lucha y los males constantes de la condición humana, motivo por el cual seguimos hablando de sus creaciones y observamos tanto sus sueños como sus advertencias en nuestra ciencia y nuestros científicos.

Turing no fue un «científico loco» en el sentido del libro de Shelley. Bletchley surgió de una colaboración, y sus esfuerzos posteriores por construir la primera computadora electrónica y universal del mundo también se debieron al trabajo en equipo. En Estados Unidos, fue testigo del talento de Von Neumann, que

acabaría ganándole de mano, armado como estaba con un genio poco habitual y el amplio grupo de recursos científicos y financieros que tenía a su disposición. Turing y Von Neumann fueron aventureros científicos, pero trabajaron en un entorno que suplantó y bañó su genio con otros talentos.

Aun así, lo más probable es que Turing tuviera algo parecido a una idea prometeica dándole vueltas por la cabeza cuando pensó en la IA, que (recordemos) según él podría mantener conversaciones genuinas hacia el año 2000. En su momento conoció a bastantes genios prometeicos, hombres que llegaron más allá del alcance habitual de los seres humanos: Einstein, el brillante lógico Kurt Gödel y, por supuesto, Von Neumann. En cuanto el formalismo de la computadora —la Máquina homónima de Turing— estuviera «ahí fuera» para que la ciencia la usara, algún científico, quizá uno que trabajara en un entorno como Bletchley, podría desvelar los secretos de la mente humana y escribirlos en código. Y es que los científicos no creen en vaguedades como la «evolución de la ciencia» más que como una capa de glaseado, como un telón de fondo. Creen de verdad en el genio científico. Lo cierto es que todos ellos se encuentran poseídos por Prometeo, por lo que las personas innovadoras pueden soñar y conseguir.

Mientras que la carrera de la IA se golpea contra una valla tras otra, el mito prometeico de la innovación extraordinaria por parte de algunos individuos está desapareciendo de los arquetipos culturales de la investigación y de la cultura más amplia. En vez de él, nos encontramos con el desarrollo pasivo de una mitología sobre la IA que crece a la par que la convicción de que la potencialidad humana se reduce.

Críticos reflexivos como Jaron Lanier han dado voz al problema central: « Deberíamos, por el contrario, intentar inspirar el fenómeno de la inteligencia humana».[1] Pero ya no quedan héroes. En su lugar tenemos «colmenas».

DE COLMENAS Y MÁQUINAS

Desplazar el epicentro de la inteligencia de las personas a las máquinas es un gambito —término ajedrecístico que implica el sacrificio de una o más piezas

para obtener una mejor posición— que de manera inevitable presenta consecuencias para la cultura humana. Si las pruebas científicas y empíricas fueran ineludibles —imaginemos que llegaran unos extraterrestres superinteligentes y no tardaran en revelarse más listos que nadie y se hicieran con el poder—, estaríamos obligados a aceptar ese gambito. A falta de tales pruebas, es una estratagema que menoscaba la cultura de la innovación y el progreso. ¿Por qué sacrificar nuestra fe en la innovación humana cuando no tenemos por qué hacerlo?

Irónicamente, se trata de una estratagema conservadora; en un momento en que se cree que los móviles están evolucionando hacia la superinteligencia, la invención radical se vuelve innecesaria. Mantenemos en su lugar diseños e ideas que benefician al statu quo a la vez que no dejamos de hablar de un «progreso» desatado. La inteligencia humana se vuelve colectiva, como una colmena de abejas, o peor, como la mente colmena del colectivo Borg de Star Trek, siempre organizado por alguien invisible entre bastidores. Básicamente, en esta mitología, la mente humana se convierte en una versión obsoleta de las máquinas del porvenir.

Pero, tal y como hemos visto, nos faltan motivos científicos para creer en todo ello, así que no deberíamos jugar a la mitología en la vida real. Lo que deberíamos hacer es construir una tecnología que nos permita ampliar una frontera de nuestra propia elección.

Eso fue lo que creímos estar haciendo durante la primera década del nuevo siglo.

EL AUGE DE LAS MÁQUINAS (QUE FUE EL AUGE DE LA GENTE)

Cuando la «Web 2.0» irrumpió en escena con un montón de nuevas tecnologías de «contenido generado por los usuarios», como los wikis y los blogs, numerosos críticos culturales y tecnológicos asumieron que nos encontrábamos en medio de una explosión de potencial humano, una nueva era de posibilidades. En 2005, la IA seguía curándose las heridas provocadas por la última burbuja «puntocom» de 2000, y el aprendizaje automático y el big data no estaban para

mucho bombo publicitario. Los blogueros ciudadanos, sí. Hacia 2005 emergió toda una escuela de pensamiento que veía la red, y en especial las tecnologías de la Web 2.0, como una nueva imprenta destinada a liberar la inteligencia y la creatividad de la humanidad. La red prometió no solo volvernos más listos e informados, sino darnos la capacidad de colaborar con mayor efectividad, construyendo pirámides digitales contemporáneas y transformando la ciencia y la cultura. No obstante, en el momento de escribir estas líneas, en 2020, las ideas originales de la Web 2.0 ya han desaparecido. De hecho, parecen francamente surrealistas.

Clay Shirky, escritor y consultor que hoy día ejerce de profesor en el programa de telecomunicaciones interactivas de la universidad de Nueva York, escribió en su momento best sellers de la Web 2.0 como Here Comes Everybody [«Aquí vienen todos»] o Excedente cognitivo: Creatividad y generosidad en la era conectada que auguraban el ascenso de un ciudadano con conciencia social y superinformado, una nueva persona.[2] Aunque un tanto patriótico, su mensaje seguía sonando fuerte y claro: los moradores de la red estaban a punto de reescribir los manuales y de liberar al mundo de esos pesados «guardianes», como la prensa y los medios tradicionales, que controlaban de manera injusta la producción y el flujo de las noticias y el conocimiento. «El poder para la gente» fue el tropo a mediados de la década de 2000, un meme que se copió y se extendió de manera interminable por blogs, comentarios y estanterías (y libros electrónicos).

Yochai Benkler, profesor de derecho empresarial en la universidad de Harvard, proclamó en 2006, en su muy leído La riqueza de las redes: Cómo la producción social transforma los mercados y la libertad, que se acercaba una nueva era, una especie de revolución en la que amplios números de personas conectadas abordarían proyectos de colaboración online, todo ello por el bien común, sin que hiciera falta mandarles cheques.[3] Wikipedia pareció respaldar su tesis, en cuanto ejemplo de producción colaborativa sin que hubiera expectativas de recompensa económica. El director de Wired, Kevin Kelly (y otros), se refirió más tarde a la oda de Benkler sobre la colaboración en línea como la mente colmena, un guiño a la inteligencia social de las abejas, sin un ápice de ironía o burla. El propio Benkler utilizó a modo de prefacio para su grito de guerra dotado de seriedad académica hacia el mundo de la Web 2.0 una cita de John Stuart Mill: «La naturaleza humana no es una máquina que se pueda construir siguiendo un modelo, de modo que realice exactamente el trabajo que se le ha prescrito, sino un árbol que necesita crecer y desarrollarse por todos lados, según

la tendencia de las fuerzas internas que hacen de él una cosa viva».[4]

Es una cita excelente. Pero, hoy en día, las palabras de Mill transmiten la sensación de ser una quimera estrafalaria, en gran medida porque su centro de gravedad apunta a la persona en vez de a la máquina.

Las ideas de Shirky también han pasado a transmitir una sensación ingenua y extravagante. El excedente cognitivo plasma la percepción de que, al conectarse todo el mundo, la gente podrá abandonar o limitar actividades que adormecen la mente como ver una comedia televisiva. En la era de internet hay un excedente de poder —pensamiento— cognitivo al que podemos dar un buen uso, provocando una revolución social en la Primavera Árabe o inventando curas contra el cáncer. El libro precursor de Shirky, Here Comes Everybody [«Aquí estamos todos»], estaba lleno de anécdotas sobre gente común y corriente que ayudaba a la policía a capturar delincuentes usando la tecnología móvil.[5]

Por supuesto que aún podemos contribuir a ello con nuestros móviles, pero el diagrama de Venn de su uso cotidiano ya no apunta hacia un futuro donde el potencial humano se hace realidad. De hecho, está claro que la revolución intelectual que se profetizó a mediados de la década de 2000 no ocurrió nunca. (La «mente colmena» ni tan solo nos dio Wikipedia, la mayor parte de cuya redacción real la llevan a cabo expertos singulares, mientras que el resto de los colaboradores realizan tareas más prosaicas de edición.) El optimismo de Shirky y compañía sobre el crecimiento humano se metamorfoseó con bastante rapidez en una visión del mundo que ve a los seres humanos como piñones de una máquina gigante. Con el tiempo, es la máquina misma —la cadena, el sistema— la que se convierte en el foco. De manera predecible, las mentes colmena acabaron promoviendo un nuevo escepticismo acerca de la inteligencia humana. La idea alimenta la mitología popular sobre el ascenso de las máquinas. Los superordenadores se han convertido en «cerebros gigantes».

Si tuviéramos que elegir el año en el que el «potencial humano» murió, como lo haría un meme, 2008 sería el favorito. El «big data» mismo entró en el diccionario. Chris Anderson, de Wired, publicó el provocativo artículo según el cual el big data iba a reemplazar la teoría científica —una sugerencia no demasiado sutil de que se podía externalizar la innovación humana y dejarla en manos de la informática.

Y, en 2008, la IA ya se había reinventado bajo su aspecto moderno de ciencia

basada en datos. En retrospectiva, la trayectoria parece evidente: desde los blogueros ciudadanos —individuales— que forjaron un nuevo futuro humano hasta las mentes colmena y los zumbidos con los que se celebró la elaboración de enciclopedias, para desembocar en el big data y en una IA que reemplace el pensamiento humano y que incluso nos libere de las molestas teorías científicas. «La naturaleza humana no es una máquina», había dicho Benkler, citando a Mill, tan solo dos años antes. Asombrosamente, las esperanzas que Benkler depositó en una humanidad conectada se vieron subsumidas por la mitología sobre la IA —una revolución de las máquinas—, que en la actualidad reemplaza e ignora la creatividad humana. Como buena parte del debate actual sobre la IA, la transformación parece haber sido motivada y concebida de manera irreflexiva.

Los científicos y otros miembros de la intelligentsia acabaron señalando que la ciencia sin teoría carece de sentido, puesto que los «modelos» o marcos teóricos anteceden al análisis del big data y prestan al aprendizaje automático algo concreto que hacer, que analizar. Pero, para 2010, el Zeitgeist de la Web 2.0 de mediados de la década de 2000 se había distanciado de manera abrupta del «poder para la gente».

Dos años después, en 2012, los sistemas de aprendizaje profundo arrasaron a sus competidores en los conocidos concursos de ImageNet que usaban conjuntos de fotos de Flickr, y no tardaron en mostrarse prometedores a extremos fantásticos con otros problemas de los consumidores como el reconocimiento de voz y la personalización de contenidos —problemas que empresas como Facebook (que ese año comenzó a cotizar en bolsa por el valor de más de dieciséis mil millones de dólares) y Google necesitaban resolver para vender anuncios y recomendar contenidos a su legión de usuarios—. Facebook, Amazon, Google y otros gigantes tecnológicos se apresuraron a abrazar la IA del big data y todo el mundo se olvidó pronto de los blogueros ciudadanos. La intelligentsia comenzó a ensalzar una IA en ciernes que publicaría blogs y redactaría las noticias por nosotros. A continuación, nos reemplazaría. En retrospectiva, la preocupación que Lanier expresó en 2010 con su «You Are Not a Gadget» [«No eres ningún dispositivo»] fue clarividente, pero llegó demasiado tarde: «Una nueva generación ha alcanzado la mayoría de edad con menos expectativas de lo que puede ser cada persona, y de aquello en lo que puede llegar a convertirse».[6]

Es posible que, hoy en día, el concepto mismo de una mente colmena parezca un poco pintoresco, aunque solo sea por la razón también deprimente de que las grandes ideas sobre el potencial humano han retrocedido. En 2005, Google

seguía siendo un prodigio, un ejemplo maravilloso de innovación humana. En la actualidad, el ubicuo motor de búsquedas gigante es como el llavero que llevamos en el bolsillo. Incluso hemos dejado de reparar en él. Menos de una década después de que James Surowiecki publicara en 2005 su exitoso Cien mejor que uno: La sabiduría de la multitud o por qué la mayoría siempre es más inteligente que la minoría, la idea de que la gente mostrara una sabiduría colectiva —o sabiduría de cualquier tipo— en Twitter y otras redes sociales parecía ya risible.[7]

Resulta revelador que la mitología alrededor de la IA no haya sido ridiculizada también, y que parezca encontrarse más en auge que nunca.

Capítulo 16

La mitología de la IA invade la neurociencia

Por mucho que ya no se invoque con la seriedad de antaño, la «mente colmena» continúa formando parte de nuestro vocabulario. Pero una ramificación ha aparecido en el más inesperado de los lugares: la ciencia misma. Colmenas para las mentes y, a continuación, «enjambres» para los descubrimientos científicos.

Sean Hill, antiguo director del Centro de Coordinación Internacional de Neuroinformática (INCF en sus siglas inglesas), forma parte de una iniciativa conjunta con el nombre de Proyecto Cerebro Humano (HBP en inglés). En el texto que publicó en la antología de 2015 The Future of the Brain [«El futuro del cerebro»], Hill anticipa que el futuro de la ciencia se basará en los esfuerzos de colaboración a gran escala, donde los científicos individuales se entenderán mejor como partes de un enjambre: «Uno de los objetivos del Proyecto Cerebro Humano consiste en desencadenar y facilitar una nueva ola de colaboración global en neurociencia ... Si se tiene éxito a la hora de involucrar a la comunidad, la idea es que haya enjambres de científicos que ataquen juntos los desafíos principales de la comprensión del cerebro y sus trastornos —en un entorno donde a cada individuo se le acrediten sus contribuciones».[1]

Este revoltijo de ideas, desde la «colaboración global», que suena prometedora, hasta los «enjambres de científicos», evoca una metáfora deflacionista y absurda sobre las contribuciones individuales de los científicos (invalidando de manera efectiva el descubrimiento individual mismo), para acabar desembocando en una perogrullada sobre que «a cada individuo se le acrediten sus contribuciones».

Es posible que Hill, uno de los actores principales del ahora infame Proyecto Cerebro Humano que se está llevando a cabo en Europa, tuviera un mal día. Pero Henry Markram, en su momento director de ese mismo proyecto, es otro defensor de la visión que Hill tiene de la ciencia, y argumenta que los genios como Albert Einstein ya no son necesarios: «Nos vemos entorpecidos por la convicción general de que necesitamos a un Einstein para que nos explique cómo funciona el cerebro. Lo que necesitamos en realidad es dejar de lado

nuestros egos y crear un nuevo tipo de neurociencia colectiva».[2] Pero la promoción que realizó de la neurociencia colectiva formó parte, según sabemos ahora, de su propia visión mitológica para crear un cerebro informático superinteligente usando a otros científicos como recursos con los que avanzar por un sendero definitivo, pero insensato.

NO HACE FALTA QUE LAS PERSONAS SE POSTULEN

La retórica sobre la ciencia como enjambre, igual que la conversación sobre las mentes colmena, conduce de manera inevitable a una visión del mundo centrada en los ordenadores, donde el potencial humano se rebaja en beneficio del ascenso de las máquinas. La ciencia sigue una cultura online, desde las ideas humanas hasta las nuevas tecnologías, y eso ha llevado a que las principales compañías tecnológicas consoliden su poder y a que el ritmo de la innovación sufra un estancamiento general.

Los futurólogos de la Web 2.0 lo saben bien. Las plataformas de «contenido generado por los usuarios» comenzaron por alumbrar nuevas visiones sobre las posibilidades humanas. A continuación, mientras la tecnología maduraba, llegaron quimeras donde la gente se conectaba en un inmenso esfuerzo colaborativo. Y finalmente la mitología sobre el carácter inevitable de la IA, que margina a la gente en sus narrativas sobre el futuro de las máquinas.

Es la misma tendencia que se está dando ahora mismo en la investigación de base. La ciencia, antaño un triunfo de la inteligencia humana, parece ahora encaminarse hacia una ciénaga retórica sobre el poder del big data y de los nuevos métodos informáticos, donde el científico ha pasado a desempeñar el papel de un técnico que en esencia se dedica a comprobar teorías ya existentes en los superordenadores Blue Gene de IBM.

Pero los ordenadores no tienen capacidad de percepción. Las personas, sí. Y los esfuerzos colaborativos solo resultan efectivos cuando se valora a los individuos. Alguien deberá tener una idea. En Bletchley, Turing era consciente de eso —o lo aprendió—, pero la lección se ha perdido durante las décadas que han transcurrido desde entonces. La tecnología —o, mejor dicho, la tecnología de la

IA— nos está atrayendo hacia su interior. Es pasmoso, pero estamos denigrando a Einstein para dejar sitio a una maquinaria parlante.

Los proyectos de neurociencia son un caso desafortunado pero ilustrativo.

EL PROYECTO CEREBRO HUMANO

El Proyecto Cerebro Humano se lanzó de manera oficial en octubre de 2013, con un premio de mil trescientos millones de dólares a lo largo de diez años que corrió a cargo de la Unión Europea, una suma inmensa para la investigación exploratoria en neurociencia. En un primer momento, el proyecto implicó a más de ciento cincuenta instituciones de todo el mundo y lo dirigió el doctor Henry Markram, neurocientífico del Instituto Federal Suizo de Tecnología en Lausana. Markram es conocido por el proyecto Blue Brain, el ambicioso intento de crear a escala una columna neocortical entera del cerebro de una rata a través de una simulación informática realizada en un superordenador Blue Gene de IBM.

Los objetivos del Proyecto Cerebro Humano expanden el alcance de Blue Brain hasta incluir nada más y nada menos que la simulación completa por ordenador de la totalidad del cerebro humano, objetivo que, según lo que Markram anunció durante una charla TED de 2009, se alcanzaría hacia el final de la década —aunque muchos otros neurocientíficos no estuvieron de acuerdo—. Al igual que las manifestaciones futuristas acerca de la IA, la predicción de Markram se reveló errada —muy errada— y, por fortuna para la ciencia, el fracaso de su pronóstico no fue ignorado por completo. En un artículo para The Atlantic en 2019, Ed York señaló de manera sucinta lo que otros neurocientíficos llevaban avanzando todo ese tiempo: «Han transcurrido diez años. No lo ha conseguido».[3]

Tras ser acusado de mala gestión, el doctor Markram renunció al cargo de director del proyecto a los dos años de su lanzamiento. Unos años después, el Proyecto Cerebro Humano se renovó sin la menor vergüenza como un simple proyecto de software que proporcionaría métodos y herramientas a los científicos humanos para que realicen sus investigaciones en curso y otras potencialmente importantes.

La idea del enjambre que realiza investigaciones científicas fundamentales es un error flagrante, y deberíamos denunciar la ya familiar tendencia —consciente o inconsciente— de subestimar la participación humana y reenfocar nuestros esfuerzos en los superordenadores y el big data, extrapolando a partir de la tecnología existente, y al mismo tiempo debemos enfrentarnos al bombo publicitario que se le ha dado a la IA.

También resulta inquietante que las ideas de Markram y de Hill acerca del futuro de la neurociencia sean ejemplos públicos de una tendencia mundial de científicos que pretenden dar empuje a la ciencia desde la informática y no desde las ideas. Markram y el Proyecto Cerebro Humano quizá representen el ejemplo más indignante de esa mitología sobre la inteligencia que ha «emergido» de la megaciencia y que tiene a los superordenadores como pieza central; pero, aunque de manera menos pública, otros proyectos han cometido los mismos errores.

Por ejemplo, pisándole los talones al premio del Proyecto Cerebro Humano, la administración Obama anunció un esfuerzo de megaciencia igual de ambicioso, el proyecto de Investigación del Cerebro a Través del Avance de las Neurotecnologías Innovativas (BRAIN*h en sus siglas en inglés), en el que invirtió cien millones de dólares durante el año fiscal de 2014, con una previsión de gastos de trescientos millones a lo largo de una década. La iniciativa BRAIN se centra en el desarrollo de tecnologías que puedan copiar circuitos neuronales y otras zonas funcionales del cerebro que comprendan múltiples neuronas individuales. Y también se han puesto en marcha proyectos de simulación cerebral más pequeños, pero importantes, como el Atlas Cerebral Allen del Instituto Allen de Neurociencia de Seattle.

Esos proyectos prometen una comprensión completa del cerebro. Markram, para comenzar, lleva años diciendo en público que planea dar cuerpo a su simulación por superordenador utilizando la robótica para crear así la primera inteligencia no biológica del mundo. Por el camino, los proyectos del megacerebro satisfacen a las agencias gubernamentales prometiéndoles más beneficios prácticos, como el conocimiento sobre las causas del Alzheimer y otras enfermedades neurodegenerativas. El presidente George H. W. Bush declaró en su día que la de 1990 fue «la década del cerebro». Parece más bien que es nuestra década la que merecería ese título. Que se haya realizado algún avance fundamental y real es otra historia.

Aunque las iniciativas del Proyecto del Cerebro Humano y BRAIN son con claridad proyectos de megaciencia —con objetivos de perfil alto, gestión vertical, bien financiados y con un foco en la ingeniería antes que en la teoría—, también han sido catalogadas como proyectos de AI. En concreto, las empresas neurocientíficas de importancia se están promocionando, casi de manera universal, como proyectos de big data. Los datos que se requieren para la realización de estos proyectos sin duda hacen que el término sea merecido. La Fundación Kavli ha señalado que la iniciativa BRAIN debe sobrevivir a una «avalancha de datos»: «Medir tan solo una fracción de las neuronas del cerebro de un ratón podría generar casi la misma cantidad de datos que los veintisiete kilómetros de largo del Gran Colisionador de Hadrones o los más avanzados observatorios astronómicos».[4] Kavli subraya una cuestión que aparece de manera prominente en la bibliografía de ambos proyectos de Gran Cerebro: el matrimonio entre la IA basada en datos y la neurociencia representa tanto un desafío de información tecnológica como una oportunidad enorme, ya que se considera que la capacidad de manipular mayor cantidad de datos sobre el cerebro llevará al éxito de la investigación.

En efecto, el big data es el plato fuerte del debate sobre los proyectos de Gran Cerebro que se están realizando. El propio Markram, por ejemplo, insiste en que el Proyecto Cerebro Humano trata sobre la integración de datos (¿y sobre la neurociencia no?). Y Amye Kenall escribe en BioMed Central, comentando la búsqueda por parte del Proyecto Cerebro Humano de un «nuevo superordenador», que «las neurociencias van a superar con facilidad a la genómica como la ciencia con un mayor uso intensivo de datos». Tal y como señala Kenall, los superordenadores actuales se ejecutan a una «petaescala», mientras que se espera que el Proyecto Cerebro Humano requiera de unos recursos informáticos de «exaescala», así que el proyecto, junto con la investigación en neurociencia, también está financiando el desarrollo del primer ordenador a nivel exaescala.[5]

Esa línea borrosa entre la informática y la tecnología de la información y la investigación en neurociencia es típica de ambos proyectos. Dados los objetivos declarados de las iniciativas del Gran Cerebro, ese foco sobre los conceptos y técnicas de la inteligencia artificial resulta, por supuesto, necesario. Ambos proyectos se enfrentan a lo que se conoce como «el problema del mapeo cerebral», un problema con una complejidad informática inherente. Rebecca Golden explicó el problema del mapeo cerebral de esta manera en un texto para el Proyecto de Alfabetización Genética:

Se calcula que el cerebro humano tiene aproximadamente 86.000 millones de neuronas (8,6 x 1010), y es posible que cada neurona tenga decenas de miles de conexiones sinápticas; pequeños lugares de conversación en los que las neuronas intercambian información. En total, con toda probabilidad habrá más de cien billones de sinapsis neuronales —de modo que, para registrar una simple información binaria sobre las sinapsis, como si una de ellas se dispara en una ventana temporal o no, haría falta un ordenador de cien terabytes—. Para almacenar esa información tan simple durante cada segundo a lo largo de un día y sobre una sola persona harían falta más de cien mil terabytes, o cien petabytes. Hoy en día, los superordenadores cuentan con unos diez petabytes. Y este cálculo rápido no tiene en cuenta los cambios de conectividad y posicionamiento que se dan en esas sinapsis con el paso del tiempo. Contar la manera en que cambian las conexiones tras una noche en la que se ha dormido bien o tras una clase de matemáticas conduce a una cifra gigante (y a muchos más bytes que los 1080 átomos que se cree que hay en el universo). El problema de cableado parece de una magnitud irresoluble.[6]

Por supuesto, Markram y otros investigadores son conscientes de la aparente insolubilidad del problema del mapeo en la neurociencia, y ahí radica su lealtad crónica al big data de la IA como fuerza motriz. Si han de surgir nuevos conocimientos sobre la «integración de datos», tal y como lo llama él, los principios que vinculan a las neuronas en forma de circuitos y de unidades funcionales más amplias (los circuitos MESA) reducirán el problema del mapeo y simplificarán la complejidad informática a la que se enfrentan los investigadores en la actualidad.

En otras palabras, el big data y la IA suministrarán las piezas que le faltan a la teoría de la neurociencia misma. Desde esta visión, la tecnología de la información, que los críticos de los proyectos del Gran Cerebro consideraban una distracción, forma parte en realidad del argumento por el que esos proyectos acabarán teniendo éxito: la tecnología, la IA misma, llenará los huecos allí donde la humanidad ha fallado hasta el momento.

BIG DATA, DE NUEVO

Los proyectos de datos cerebrales, y en particular el Proyecto del Cerebro Humano, representaron intentos audaces por potenciar un panorama general de la ciencia abordando temas capitales de la naturaleza del pensamiento humano. En efecto, Markram, que no se ha dejado intimidar por el fracaso del proyecto, ha proclamado sin la menor vergüenza que los proyectos de datos cerebrales son una ruta hacia la IA. Cree que la IA y la neurociencia resolverán el misterio de la inteligencia humana, y quizá también el de nuestra conciencia. En las numerosas entrevistas publicadas que ha concedido a lo largo de los años, Markram ha manifestado que piensa crear modelos de neuronas sirviéndose de datos sobre el cerebro facilitados por proyectos de todo el mundo para descubrir los «principios estadísticos» sobre los que se afianza la actividad neuronal (al nivel de los intercambios iónicos) y para relacionar unidades funcionales de neuronas del cerebro humano cada vez más amplias hasta que surja un mapa completo de estas.

Según él, ese mapa tendrá la capacidad de exhibir conductas parecidas a la humana. Los futurólogos como Ray Kurzweil, junto con una bandada de creadores de mitos relacionados con la IA, también insisten en esa idea. Por decirlo de otro modo, llegaremos a una comprensión tal de los principios del pensamiento inteligente que estos se podrán reducir a una labor de ingeniería que programaremos en la robótica y en los sistemas de AI. De la mano de una serie de avances capitales en neurociencia, asistiremos al lanzamiento de una nueva era de la IA. Y, una vez más, tendremos éxito allí donde las generaciones precedentes fracasaron gracias a la IA: el acceso a volúmenes de datos, a la integración de esos datos y a las plataformas de análisis nos permitirá descubrir principios y teorías allí donde antes nos inundaban las diversas investigaciones.

En otras palabras, la IA es lo que lleva a que las iniciativas del Gran Cerebro parezcan nuevas y diferentes en comparación con lo que había antes: no son los hallazgos de la teoría neurocientífica, sino la capacidad informática y sus datos. Pero se trata de un absoluto disparate, tal y como demuestra con claridad el fracaso del Proyecto de Cerebro Humano. Igual que cuando se compara la inducción con la abducción, la tecnología es un derivado de la teoría, y la tecnología de la información en especial está reemplazando a la innovación mientras los investigadores pregonan su innovación e implementan marcos de

investigación tecnológica utilizando ideas existentes.

Probablemente sea cierto —o cuando menos es razonable asumirlo— que el carácter inadecuado del conocimiento neurocientífico es una de las razones clave por las que no disponemos de mejores teorías sobre la naturaleza de nuestra mente. En especial, una mejor comprensión de los principios de la cognición humana podría dar lugar a una IA que se dedicara a desbloquear el misterio de la inteligencia —uno de los objetivos declarados de Markram—. Sin embargo, las simulaciones del cerebro a gran escala (los enfoques de datos cerebrales) parecen haberlo entendido todo del revés a la hora de buscar ese conocimiento. A falta de una teoría sobre la manera en que se conduce el cerebro —cómo pensamos y sentimos y percibimos—, el conocimiento existente, sobre las neuronas y unidades funcionales expresadas como circuitos en simulaciones por ordenador, nos hace albergar la esperanza de que los ingredientes que le faltan a la cognición vayan a emerger de algún modo de los volúmenes de datos recolectados en esos niveles inferiores.

Se trata, por supuesto, de una presunción fundamental de la mitología acerca de la IA misma: que las percepciones, teorías e hipótesis que se desconocen y que de hecho son incognoscibles a pequeña escala brotarán de golpe cuando hayamos recogido y analizado los datos suficientes, a través del aprendizaje automático y de otros enfoques inductivos.

Los afanes relacionados con los datos cerebrales, tanto el del Proyecto Cerebro Humano como el de la iniciativa BRAIN en Estados Unidos, respaldan esta dudosa idea con la esperanza de que la neurociencia vaya a seguir en general el ejemplo del Proyecto Genoma Humano y demostrar así que la ciencia se puede reducir a una labor de ingeniería (y los científicos, a enjambres de ayudantes).

Pero debería señalarse que el Proyecto Genoma Humano tenía unos objetivos bien definidos, que obviaban los principales desafíos teóricos —se trató de un proyecto de ingeniería desde el primer momento—. Por suerte, no sucede así con todos los proyectos de datos cerebrales. Los investigadores en neurociencia —desde Eric Kandel, en la universidad de Columbia, hasta Christof Koch, antes en Cat-Tech y ahora en el Instituto Allen de Neurociencia, e incluso Markram— admiten que en nuestra imagen del cerebro faltan piezas importantes de la teoría, en especial en los niveles más altos de funcionalidad. La respuesta, que en su mayor grado de virulencia viene firmada por Markram y el aún ambicioso Proyecto del Cerebro Humano, dice que el big data ocupará el lugar de las

teorías que no existen y que los científicos están perdiendo el tiempo al investigar a la manera tradicional —con objetivos pequeños y bien definidos sobre zonas problemáticas concretas—. En la era de la IA, al parecer, no podemos esperar a que la teoría surja del descubrimiento y la experimentación. Tenemos que depositar nuestra fe en la supremacía de la inteligencia informática sobre la humana —asombrosamente, haciendo frente al misterio teórico sin solucionar que es el imbuir a los ordenadores una inteligencia flexible.

Esta manera de pensar constituye un error y, cuando todo se calme, lo más probable es que acabe revelándose muy cara.

De hecho, pese a su retórica constante sobre el progreso, los proyectos de datos cerebrales manifiestan una naturaleza conservadora en lo referente a la investigación neurocientífica misma. Los investigadores del Proyecto Cerebro Humano, por ejemplo, parecen estar encantados de usar la investigación existente en neurociencia como base para sus planes de «integración de datos», creyendo erróneamente que los datos por sí solos proporcionarán respuestas a medida que la complejidad de la simulación vaya creciendo —uno de los motivos de arrogancia originales del big data—. Por tanto, el proyecto, junto con otras iniciativas parecidas basadas en la IA, socava de manera efectiva el proceso del descubrimiento científico al poner el énfasis sobre esas simulaciones cada vez mayores de un conocimiento experimental ya existente.

Que los proyectos de perfil alto sobre datos cerebrales —como el Proyecto Cerebro Humano europeo y, ahora, la iniciativa BRAIN en Estados Unidos— hayan sido capaces de convencer a los organismos financieros y, en gran medida, al público ignorante de que ese saber existente, simulado en plataformas de supercomputación que procesan grandes volúmenes de datos, constituye un avance científico en lugar de un proyecto de ingeniería conservador, pone de manifiesto la confusión que existe sobre el papel de la IA en la ciencia en general.

Por ejemplo, una técnica que en un primer momento impulsó Markram, y que fue adoptada por el Proyecto Cerebro Humano, se ha extendido ahora, bajo aspectos diferentes, a numerosos proyectos neurocientíficos de importancia. Se conoce como «neurociencia predictiva» y es un método en el que los investigadores simulan las conexiones interneuronales. Los enlaces sinápticos aún desconocidos se determinan a partir de enlaces conocidos mediante técnicas de aprendizaje automático inductivo. Se usan tanto las redes neuronales

tradicionales como redes más poderosas del aprendizaje profundo. En un primer momento, Markram demostró que el aprendizaje automático puede predecir con acierto algunas conexiones hasta entonces desconocidas en las columnas corticales del cerebro de la rata. Un análisis en el que se usó un valor estadístico estándar de F brindó una exactitud de cerca del 80 % para ese método.

Aunque representara un avance de cara a aplicar el aprendizaje automático a los conjuntos de datos biológicos, el método cuenta con un error promedio de dos sobre diez, lo cual presenta implicaciones inquietantes para cualquier estrategia que pretenda realizar una ingeniería inversa del cerebro humano. Pero, al margen de esa preocupación, es la lógica superficial que hay tras esos enfoques lo que anuncia problemas de mayor calado para la institución de la ciencia.

EL BOSÓN DE HIGGS

En 2012, los científicos descubrieron una pieza perdida desde hacía mucho tiempo en el modelo estándar de la física, el bosón de Higgs. Ese descubrimiento se suele atribuir a un impresionante artefacto tecnológico, el Gran Colisionador de Hadrones (LHC en sus siglas inglesas), que se extiende a través de la frontera franco-suiza. El LHC cuenta con 27 kilómetros de tuberías, con lo que es el supercolisionador más grande del mundo.

Los científicos usaron el LHC para diseñar un experimento que debía probar la existencia de una partícula responsable de la masa del universo, bautizada como bosón de Biggs en honor de Peter Higgs, el primer científico que predijo su existencia. El resultado de ese experimento de alta energía en el LHC pareció confirmar la teoría de Higgs sobre la existencia de la partícula (esta se confirmó de manera oficial en 2013).

Pero el caso del bosón de Higgs también ilustra la tendencia perversa a minimizar la teoría y defender los métodos informáticos que refuerzan el pensamiento ilusorio sobre el big data y la IA.

Es cierto que el recurso de los superordenadores resulta necesario para darle un sentido a la montaña de datos generados por el supercolisionador. En 2012, el

colisionador generó unos 25 petabytes al año; en 2018, el número se había duplicado y equivalía a unos quince millones de películas en alta definición. No hay duda de que el análisis del big data y los recursos informáticos necesarios para procesar esos volúmenes de datos otorgan a los físicos un juego de herramientas poderosas con el que explorar el mundo subatómico. Pero la historia del bosón de Higgs, a menudo publicitada como ejemplo del éxito de la megaciencia (que lo fue), es también un triunfo de la percepción teórica. El de Higgs es un caso especialmente impresionante en su teoría, no desde los macrodatos per se. En realidad, Peter Higgs había descubierto la partícula en 1964; el LHC se limitó a confirmar su existencia. Se trata de un caso práctico sobre el uso adecuado de la tecnología para suplantar la percepción humana. La lección de la confirmación de la partícula de Higgs no representa una llamada a las armas de la mitología para nuestro futuro informático. En cambio, es un recordatorio de que la IA —aquí, el big data— funciona solo cuando disponemos de una teoría previa. Sobre ese punto descansa precisamente una gran confusión y no menos problemas en potencia para la ciencia.

Por desgracia, la neurociencia, a diferencia de la física de partículas, no dispone de un marco teórico que la unifique. El desafío para los neurocientíficos, pues, consiste en defender el actual modelo basado en datos a falta de los conceptos teóricos que presten sentido y que dirijan la «avalancha de datos», en las palabras de la Fundación Kavli.

SE BUSCA: TEÓRICOS

Hace unos años, durante una reveladora entrevista en la revista Nature, el prominente neurocientífico Eric R. Kandel (director del Instituto Kavli de Neurociencia de la universidad de Columbia), Henry Markram (director del Proyecto Cerebro Humano), Paul M. Mathews (responsable de la División de Neurociencias del Imperial College de Londres, Reino Unido), Rafael Yuste (profesor de ciencias biológicas y neurociencia en la universidad de Columbia) y Christof Koch (director científico del Instituto Allen de Neurociencia en Seattle) debatieron el papel del big data, la teoría y la colaboración en el Proyecto Cerebro Humano europeo y en la iniciativa BRAIN norteamericana.[7] Ese debate dejó claro que la neurociencia se enfrenta a unos desafíos capitales, y que hay

escasez de respuestas. Mathews, por ejemplo, admitió que «tanto la iniciativa BRAIN como el Proyecto Cerebro Humano se enfrentan a un desafío fundamental: no contamos con un paradigma potente que dirija nuestra investigación. Resulta llamativo que tanto la iniciativa BRAIN como el Proyecto Cerebro Humano sean ejercicios de recolección de “big data”, de los que se espera que surjan relaciones cargadas de contenido».[8]

Sin embargo, Markram mantuvo la confianza en que la IA de big data rellene esos huecos teóricos a medida que vaya avanzando y fue explícito al respecto al manifestar que «a mayor cantidad de datos, más biológicamente precisos se volverán los modelos». En la entrevista con Nature dejó clara su visión sobre el Proyecto Cerebro Humano y sobre la neurociencia en general: «Científicamente, queremos abrir camino para una nueva forma de neurociencia acelerada en la que identifiquemos unos principios básicos que abarquen múltiples niveles de organización cerebral y los explotemos para llenar los amplios vacíos que existen en nuestro saber. Por ejemplo, podemos usar los principios sobre la manera en que se conectan las neuronas para predecir el conectoma [el diagrama del cableado del cerebro]. Las reconstrucciones hipotéticas del cerebro pueden guiar y acelerar el mapeo experimental del cerebro, y hacer que pase de ser un sueño a convertirse en una realidad práctica».[9]

La lectura de las voluminosas entrevistas de prensa de Markram evidencia que cree que la integración de datos —resultados relevantes de investigaciones procedentes de todas partes del mundo y recogidos en su plataforma técnica instalada en un superordenador— facilitará la «emergencia de teorías» a niveles cada vez más elevados de la organización cerebral. Por tanto, los modelos neuronales iniciales, que usan información sobre los canales iónicos (esto es, información de bajo nivel molecular sobre la conducta de una neurona individual), ayudarán a generar teorías sobre la manera en que operan unidades funcionales como los circuitos neuronales, lo que a su vez permitirá sugerir principios o teorías que permitan a los investigadores conectar microcircuitos a circuitos MESA y a regiones cerebrales enteras como el neocórtex.

El propio Markram se muestra muy claro sobre la manera en que funcionará todo, y desdeña también la sugerencia de que una investigación creativa por parte de equipos más pequeños pueda llegar a esas teorías de manera independiente de su método de big data. Su interés consiste en usar a los científicos y sus investigaciones como entradas en un marco tecnológico centralizado, motivación que admite abiertamente, desde la creencia de que nos

conducirá a «un nuevo tipo de neurociencia colectiva», sin Einstein.[10]

Aunque cuesta discutir ideas tan vagas y de sonido tan positivo como la de «colaboración», lo que aguarda más allá de esos comentarios honoríficos son unas manifestaciones grandilocuentes acerca del papel del big data y de la IA como motores conductores de la neurociencia del futuro. Al parecer, Markram cree que los datos mismos (y el aprendizaje automático) guiarán la construcción de teorías, a pesar de que la historia de la ciencia y la corta pero explosiva moda pasajera del big data revelan el despropósito de ese enfoque.

Ya hemos visto que los éxitos del big data promocionados en las cuentas populares (como el Big Data de Cukier y Mayer-Schönberger) provienen en gran medida de ámbitos sociales como los negocios, donde no se conoce que existan teorías sólidas que respalden el comportamiento. A falta de una teoría, la IA del big data ha sido una bendición para numerosas áreas de interés en las que, sin métodos intensivos de datos, no se podría haber realizado un avance real y predictivo. Esos ejemplos pueden resultar alentadores para los líderes empresariales y pueden incluso arrojar luz sobre zonas de interés de la cultura popular, pero por lo general resultan inadecuados e incluso desastrosos para la ciencia seria.

Ya hemos visto la manera en que ha operado la teoría en supuestos éxitos tecnológicos como el descubrimiento del bosón de Higgs. En esos casos, un marco teórico robusto posibilita un conjunto de experimentos controlados y enfocados que pueden ayudar a confirmar los resultados anticipados por la teoría. En ese sentido, la ciencia siempre se ha aliado con la experimentación, así que es difícil que el caso de Higgs resulte controvertido. Pero tampoco hace gran cosa por apuntalar las manifestaciones radicales acerca del poder de la IA del big data según las líneas de lo que necesitaría Markram para la neurociencia.

De nuevo, la ausencia de un marco teórico sólido en neurociencia hace que el enfoque centrado en los datos y en el aprendizaje automático resulte aún más desencaminado. A menor cantidad de teorías que tengan los investigadores para dirigir los métodos basados en datos, menor será la cantidad de hipótesis bien definidas que podrán comprobar, y mayor el número de esos enfoques que caigan víctimas de las debilidades ya conocidas en los métodos basados en datos.

Michael Jordan, miembro del Instituto de Ingenieros Eléctricos y Electrónicos de

Estados Unidos (IEEE en sus siglas inglesas) y profesor de ciencias informáticas con la distinción Pehong Chen en la universidad de California en Berkeley, se ha manifestado en contra de quienes ven una conexión lineal entre el big data y el pensamiento científico (idea según la cual cuantos más datos tengamos, mejor será nuestro pensamiento científico). Tratándose de una de las autoridades más respetadas del mundo en aprendizaje automático y big data, Jordan es un crítico insólito, pero predice que «la sociedad está a punto de experimentar una epidemia de falsos positivos procedente de los proyectos de big data». En sus propias palabras: «Cuando dispones de grandes cantidades de datos, tu apetito por las hipótesis tiende a crecer aún más. Y, si crece a mayor velocidad que la fuerza estadística de los datos, lo más probable es que muchas de nuestras inferencias acaben siendo falsas. Con toda probabilidad serán ruido blanco».[11]

Jordan afirma sobre cualquier conjunto de datos concreto que «puedo encontrar alguna combinación de columnas que anticipe a la perfección cualquier resultado... a partir tan solo del azar, encontraré todo tipo de combinaciones espurias de columnas porque hay un número inmenso de ellas. Así que es como tener a miles de millones de monos tecleando el ordenador. Uno de ellos acabará escribiendo Shakespeare».[12]

Jordan señala aquí un problema muy conocido de la estadística que se conoce como «sobreajuste» (lo comentaremos en breve). Tristemente para Markram y otros abogados de las iniciativas de cerebro de datos como el Proyecto Cerebro Humano, el sobreajuste resulta especialmente problemático en ausencia de una información causal o teórica sobre cualquier ámbito —es decir, en ausencia de una inteligencia—. Dice una advertencia familiar que «la correlación no implica causalidad», pero algunas manifestaciones audaces que se han realizado estos últimos años en favor de los proyectos de la IA de big data han hecho que vuelva a resultar especialmente relevante. Tal y como Marcus y Davis se vieron obligados a señalar en el New York Times, «un análisis del big data puede revelar, por ejemplo, que entre 2006 y 2011 el índice de asesinatos en Estados Unidos tuvo una correlación clara con la cuota de mercado del Internet Explorer: ambos descendieron abruptamente. Pero cuesta imaginar que haya podido haber alguna relación causal entre ambos».[13]

Es una cuestión evidente, que se puede aplicar con fuerza a las declaraciones sobredimensionadas acerca del aprendizaje automático. Las expresiones de moda han cambiado, pero la tendencia principal que viene proyectando la IA del big data como una panacea amenaza nuestro progreso en áreas fundamentales

como la de la neurociencia, pese a las afirmaciones atrevidas que han realizado entusiastas como Markram y otros. La moraleja de esta historia es que el mito acaba teniendo consecuencias prácticas para nuestro futuro como seres humanos, es decir, en la ciencia de verdad.

EL SOBREAJUSTE

El estadístico Nate Silver también ha señalado el peligro inherente de sobreajustar las teorías (los modelos) a los datos, donde ese «sobreajuste» implica emparejar de manera falsa un conjunto de puntos de datos con una descripción que no contiene ninguna capacidad explicativa, porque la descripción no se puede generalizar ante otros puntos de datos nuevos y desconocidos en la distribución subyacente en cuestión. La generalización implica un alejamiento abstracto respecto a los detalles irrelevantes de los datos y el aislamiento de las relaciones genuinas en la representación del principio o teoría.

El caso más simple de una «teoría» o modelo de conjunto de puntos de datos es la interpolación lineal de un diagrama de dispersión. Cada punto de datos sobre el sistema de coordenadas podría trazarse con una línea más compleja que describa los puntos existentes, pero esa descripción sería inútil para los nuevos puntos marcados, ya que no contiene ninguna información sobre la distribución de los puntos. En cambio, una línea recta nos muestra la interpolación media o lineal de los datos desperdigados, y por tanto nos ofrece un modelo útil para predecir su comportamiento. El sobreajuste, tal y como señala Silver, transmite una falsa confianza en los datos existentes, pero se apresura a hacer trizas esa ilusión en cuanto aparecen nuevos datos —y estos no encajan en el modelo o teoría.[14]

El sobreajuste es un problema reconocido en el análisis e inferencia estadísticos, y con frecuencia se ha convertido también en motivo de grandes fracasos en importantes proyectos científicos. Aquí, una vez más, la disponibilidad de una teoría puede ayudar a los investigadores a mantenerse alejados de los modelos sobreajustados y sus correlaciones espurias. Tal y como indica Silver, varios intentos de perfil alto para predecir terremotos (usando datos históricos sobre

seísmos e información geográfica detallada sobre energías de tensión y demás fenómenos que se dan bajo la superficie de la tierra, a lo largo de las líneas de falla) fracasaron penosamente pese a encajar a la perfección con los datos existentes sobre el tema.

Igual que la línea serpenteante que explica los puntos existentes sobre un diagrama de dispersión, los modelos resultaron carecer de cualquier valor predictivo o científico. Son numerosos los fiascos relacionados con la predicción de terremotos por parte de los geólogos, según señala Silver, y su punto culminante llega con el fracaso del geofísico y matemático ruso Vladímir Keilis-Borok, quien predijo un terremoto en el desierto de Mojave para 2004 usando un modelo estadístico «elaborado y opaco» que identificaba patrones a partir de terremotos menores en regiones concretas y los generalizaba para anunciar los de mayor tamaño.

David Bowman, alumno de Keilis-Borok que hoy en día ocupa la cátedra del departamento de geología de Cal State Fullerton, admitió, en un raro episodio de humildad científica, que el modelo de su maestro estaba simplemente sobreajustado.

A continuación, Bowman contó que la predicción de terremotos se encuentra limitada por la ausencia de una comprensión teórica acerca de lo que sucede bajo la superficie de la tierra, a lo largo de las líneas de falla. A falta de una teoría genuina que usar como guía de los enfoques basados en datos o estadísticos, los modelos corren el peligro constante de captar «ruido» en vez de «señales», explica Silver.

LA NEUROCIENCIA: ¿SE NIEGA A APRENDER DE LOS ERRORES AJENOS?

Estas lecciones deberían ser especialmente pertinentes para la neurociencia actual, ya que los enfoques del big data parecen condenados a repetir los errores que se han cometido en otras áreas de la ciencia. De manera reveladora, al observarlo así, el énfasis que se ha puesto en la IA del big data no resulta especialmente nuevo ni alentador. Cabría decir que la teoría no puede eliminarse

por completo de la ciencia. Un aspecto irónico del mito de la inevitabilidad es que la teoría resulta necesaria no solo para llegar a una ciencia genuina, sino también para cumplir el sueño de una inteligencia general en la IA. La confusión y la mitología contemporáneas llevan a que sea la cola la que menea al perro.

Sin embargo, existen teorías sobre los picos de actividad de las neuronas y el papel de unidades funcionales de menor tamaño, como los circuitos neuronales, en algunas regiones del cerebro. Incluso existen teorías de alto nivel, cognitivas o de inteligencia, basadas en el funcionamiento del neocórtex humano. Lo que no existe, tal y como han señalado Markram y otros, es un marco unificador o una teoría que haga encajar estas piezas tan dispares.

El esfuerzo actual en el ámbito de los datos cerebrales está comprometido con la idea de que el progreso resulta posible gracias al aprendizaje automático y la IA de base. Cuesta tomarse esto en serio, al menos por dos motivos. Uno, las grandes teorías de la inteligencia inspiradas en el neocórtex que dan vida a buena parte de la visión de los proyectos de datos cerebrales que reproducen la mente humana en silicio son irremediablemente generales e inútiles. Las teorías mismas sirven de muy poco (es irónico) a los marcos informáticos o de ingeniería de Inteligencia Artificial, porque no nos dicen lo suficiente sobre lo que hace el cerebro mientras genera un comportamiento inteligente. Por consiguiente, las grandes las teorías existentes ya sugieren que los métodos basados en datos nos están proporcionando un conjunto de postulados mediocre y demasiado genérico para comenzar.

Benévolo, Markram sin duda cree que el progreso a partir de cero, a través de la simulación de unidades funcionales del cerebro cada vez mayores, logrará mejorar o complementar de algún modo los modelos de cognición relacionados con el neocórtex. Una conclusión más acertada diría que poner el énfasis en una IA del big data parece ser una excusa para proponer una serie de teorías vagas y descartables, donde los detalles en sí y el éxito último de la neurociencia se abandonan ante unas afirmaciones casi mitológicas sobre el poder de los conjuntos de datos amplios y la informática inductiva. Se supone que, allí donde las personas no han logrado arrojar luz sobre un ámbito complejo con una teoría comprobable, el aprendizaje automático y el big data pueden dar un paso adelante y hacer que la preocupación tradicional por dar con una teoría sólida se vuelva inoperante. Esa parece ser la lógica de los proyectos actuales del cerebro de datos.

Para ver de primera mano el estado demasiado generalista y relativamente pobre de las grandes teorías existentes, a continuación nos centraremos en un estudio sobre las hipótesis acerca de la inteligencia humana inspiradas en el neocórtex. Lejos de demostrar la vigencia de la teoría en la neurociencia, muestran una voluntad creciente por someter las teorías procedentes de esa disciplina al paradigma informático tan popular en la actualidad.

Eso sugiere que las debilidades de esas teorías hacen que no se pueda disuadir a neurocientíficos visionarios como Markram, en gran medida por culpa de la creencia predominante de que el trayecto entre que va de la IA del big data a la inteligencia general y más allá, permitirá añadir los detalles más adelante. En vez de suponer un reto para las actividades actuales de la neurociencia, los defensores del Cerebro de datos están cada vez más decididos a entregar sus misterios y debilidades a la supuesta magia de la IA de datos.

A continuación, nos centraremos en esas teorías basadas en el neocórtex.

Capítulo 17

Las teorías de la inteligencia humana basadas en el neocórtex

El científico informático, emprendedor y defensor de la neurociencia Jeff Hawkins ha presentado una popular teoría sobre la inteligencia. Famoso por haber desarrollado la Palm Pilot y como lumbrera omnipresente de Silicon Valley, Hawkins probó la temperatura de las aguas de la neurociencia (y de la inteligencia artificial) en 2004 con la publicación de On Intelligence [«Sobre la inteligencia»], un intento audaz y original por resumir los volúmenes de datos neurocientíficos sobre el pensamiento en el neocórtex a partir de un modelo jerárquico de inteligencia.[1] Desde entonces ha creado una empresa, Numenta, dedicada a desbloquear los secretos de la inteligencia en términos de computación.

Hawkins afirma que el neocórtex recibe los impulsos de nuestros sentidos y los «descodifica» en capas jerárquicas, donde cada capa superior realiza sus predicciones a partir de los datos que le proporcionan las capas inferiores, y así hasta que se llega a lo alto de la jerarquía y una teoría predictiva general se sintetiza desde el rendimiento de estas capas inferiores.

Su teoría da sentido a algunos datos empíricos, como la diferencia que existe en nuestras respuestas según los diferentes tipos de inputs que recibamos. En los problemas predictivos más «sencillos», la propagación por la jerarquía del neocórtex termina antes (porque la respuesta ya está disponible), mientras que, para los problemas de mayor dificultad, el córtex continúa procesando y conduciendo la entrada neural hacia capas más elevadas, potentes y globalmente sensibles. A continuación, o bien la solución está disponible, o bien se regresa a las capas inferiores hasta que llegamos a una predicción coherente basada en el input original.

LA TEORÍA BASADA EN EL RECONOCIMIENTO DE

PATRONES JERÁRQUICOS DE KURZWEIL

La naturaleza jerárquica del neocórtex también ha sido señalada por Ray Kurzweil, quien escribió en su obra de 2012 Cómo crear una mente que «el neocórtex es responsable de lidiar con nuestra capacidad para tratar con los patrones y para hacerlo de manera jerárquica. Los animales que no disponen de neocórtex (básicamente los que no son mamíferos) son en gran medida incapaces de entender las jerarquías».[2] Kurzweil reconoce que su propia teoría jerárquica, que bautizó como «Teoría de la mente humana basada en el reconocimiento de patrones» (PRTM en sus siglas inglesas), es una continuación de algunos hallazgos aceptados por la neurociencia sobre la estructura y las funciones del neocórtex humano, y da crédito a la influencia de progenitores como la teoría jerárquica del propio Hawkins.

En efecto, la estructura jerárquica del neocórtex es un aspecto bien fundamentado por la neurociencia. La organización columnada del neocórtex fue descubierta por el neurocientífico norteamericano Vernon Mountcastle en 1957. Mountcastle señaló que el neocórtex —una capa de 2,5 milímetros de grosor, compuesta por fibras nerviosas, que se extiende sobre todo el cerebro— estaba compuesto por columnas de neuronas, y que en apariencia todas eran idénticas entre sí. En el neocórtex humano hay cerca de medio millón de columnas de ese tipo, y cada una de ellas contiene unas sesenta mil neuronas.

Kurzweil lanzó una hipótesis según la cual cada columna cortical contiene lo que él denomina reconocedores de patrones, que contienen unas cien neuronas, lo que sumaría cerca de trescientos millones de reconocedores de patrones en el neocórtex humano. Igual que Hawkins, Kurzweil considera que esos hipotéticos reconocedores se organizan en jerarquías que son en esencia responsables de las capacidades únicas del pensamiento humano.

Es una hipótesis interesante. No obstante, tanto Hawkins como Kurzweil cometen el error de creer que la inteligencia humana es simple.

En estas mismas páginas he señalado la naturaleza superficial de ese tipo de teorías sobre la inteligencia humana, y no he sido el único que ha reparado en el carácter general e inútil de esos esfuerzos en pos de una teoría fundamental, que parecen más adecuados para postular ciertas arquitecturas de sistemas

informáticos que interesan a los autores, que para conducir a una comprensión real del cerebro humano. Tal y como ha manifestado Gary Marcus, tanto Hawkins como Kurzweil parecen ofrecer una teoría insustancial de la IA basada en unas percepciones vagas acerca del cerebro. Dice Marcus: «Ya sabemos que el cerebro tiene una estructura, pero la pregunta de verdad es qué hace toda esa estructura en términos técnicos. ¿Cómo se proyectan los mecanismos neuronales del cerebro sobre sus mecanismos cognitivos?».[3]

Marcus continua señalando que esas teorías resultan demasiado genéricas para hacer que la pelota avance en la neurociencia y en esfuerzos similares de la IA y las ciencias cognitivas: «A cierto nivel, casi cualquier cosa que haga cualquier animal se puede entender en clave de reconocimiento de patrones jerárquicos; ese es el motivo por el que la idea lleva dando vueltas desde los años cincuenta Pero la simple aseveración de que la mente reconoce por sí sola patrones jerárquicos nos dice muy poca cosa: no nos dice por qué los seres humanos son una clase de seres que usan el lenguaje (se supone que los roedores también cuentan con la capacidad para reconocer patrones jerárquicos, pero no hablan), y no nos explica el motivo por el que tantos seres humanos padecen problemas constantes de autocontrol, ni por qué somos un tipo de seres que dejan propina en restaurantes de pueblos a los que no regresaremos nunca».[4]

Irónicamente, esas teorías tan genéricas también hallan inspiración en la IA del big data de manera indirecta pero muy real. Kurzweil es famoso por su uso de métodos jerárquicos de aprendizaje automático en aplicaciones de reconocimiento de voz; trabajó en la aplicación original de Siri, que ahora pertenece a Apple, y en parte del iPhone. Los modelos de Márkov, ocultos y jerárquicos, forman parte de las técnicas de análisis de datos que se han fusionado con el big data. Y, en términos más recientes, las ahora omnipresentes redes de aprendizaje profundo se organizan en capas jerarquizadas. Todos esos métodos usan amplios conjuntos de datos como inputs para descubrir patrones en los datos, y generan un modelo binario que a continuación se puede descodificar sobre los datos invisibles.

En efecto, los métodos de aprendizaje jerárquico representan en la actualidad una tendencia casi tan importante como el big data mismo hace una década —mira el aprendizaje profundo—. Las teorías en neurociencia —en otras palabras, proyectos como el Proyecto Cerebro Humano y la iniciativa BRAIN— se están volviendo indistinguibles de los métodos de éxito en las ciencias informáticas y, concretamente, en la IA del big data.

LA TEORÍA DE LAS PIEZAS DE LEGO DE MARKRAM

Henry Markram es célebre por otra teoría demasiado general sobre el aprendizaje, que coloquialmente se conoce como «teoría cognitiva de las piezas de Lego» y que de nuevo basa su arquitectura en el resultado de investigaciones generales acerca de la naturaleza columnada y jerárquica del neocórtex humano.[5] Junto con el coautor Rodrigo Perin, Markram explica que, según esa teoría, «la adquisición de recuerdos se parece bastante a construir cosas con piezas de Lego. Cada ensamblaje equivale a una pieza de Lego que contiene algún fragmento de conocimiento elemental e innato sobre la manera en que hay que procesar, percibir y responder al mundo».[6]

De nuevo, una hipótesis interesante. De nuevo, es demasiado simple y demasiado mecánica.

En este momento nos encontramos en una posición que nos permite manifestar de manera explícita aquello que se ha vuelto evidente. No se trata solo de que los proyectos del cerebro de datos hayan abogado por el big data como método para rellenar los vacíos de nuestra comprensión sobre el cerebro —como, por ejemplo, en los principios emergentes del conectoma (conectividad sináptica)—, sino que varias teorías importantes parecen ligadas a los paradigmas de la ciencia informática de tal manera que es posible que la única dirección elocuente que puedan seguir ahora esos proyectos es la de las ideas y teorías explícitamente informáticas.

UNA INVESTIGACIÓN SIN FUTURO

Ya hemos visto que la IA basada en big data no es compatible con la emergencia de teorías. Por el contrario, sin las teorías ya existentes, la IA basada en big data cae víctima en general del sobreajuste, la saturación y la ceguera de los métodos de inducción de datos. Podemos añadir aquí que la informática basada en datos

viene produciendo hasta la fecha unas teorías bastante pobres y carentes de interés, que se encuentran sospechosamente ligadas a los enfoques de la tecnología popular contemporánea.

Por supuesto, esto no pasa solo con la IA basada en big data, sino también con la Megaciencia, tal y como los neurocientíficos comienzan a descubrir. Quizá nadie lo haya planteado mejor que Paul Mathews, durante la ya citada entrevista de 2013 con la revista Nature: «No se me ocurre ningún avance conceptual nuevo y de primer orden que se haya derivado de esos grandes esfuerzos científicos en el pasado».[7] Markram y demás nombres comprometidos con el propósito del Proyecto Cerebro Humano y el cerebro de datos han depositado generalmente sus esperanzas para el progreso de la neurociencia en la IA basada en big data, pero lo que de verdad hace falta —tal y como sugiere Mathews— son agendas dispares y de amplio espectro que alienten las hipótesis creativas y estimulen los descubrimientos. La IA basada en big data no es compatible con esos objetivos.

Antes de que se cumpliera el año del lanzamiento del Proyecto Cerebro Humano, la visión de Markram ya había provocado fuertes críticas por parte de un número creciente de neurocientíficos. En julio de 2014, más de quinientos científicos solicitaron a la Comisión Europea que realizara cambios importantes en el proyecto y plantearon sus motivos de preocupación, muchos de los cuales estuvieron relacionados con la fe del proyecto en la informática y el big data, en detrimento de la teoría necesaria y de la investigación creativa.

Irónicamente, la petición que se elevó ante la CE fue en parte una reacción a la decisión de Markram de desmantelar la división de arquitecturas cognitivas del proyecto —el equipo dedicado de manera específica a explorar cuestiones de cognición e inteligencia, en consonancia con la visión más amplia que Markram había manifestado—. A los neurocientíficos también les preocupó que el Proyecto Cerebro Humano no tuviera la intención de comprobar ninguna hipótesis específica, ni colección de hipótesis, relacionadas con el cerebro.[8]

Los neurocientíficos señalaron en la petición que el aumento de las simulaciones detalladas del cerebro no iba a conducir de manera inevitable a una mejor comprensión. Es decir, que centenares de científicos hicieron retroceder el Proyecto Cerebro Humano porque en realidad no tenía nada que ver con la investigación neurocientífica, sino que se trató de un proyecto de ingeniería de IA basada en big data. Markram no tardó en renunciar a su cargo, pero el proyecto se reestructuró hacia la ingeniería de software, un campo sin duda

menos infectado por la mitología de la IA pero, de manera deliberada, inoperante de cara a una investigación fundamental.[9]

Tal y como dijo Eric Kandel, neurocientífico de la universidad de Columbia, comentando el lanzamiento de la iniciativa BRAIN en Estados Unidos: «Sabíamos su punto final [el del Proyecto Cerebro Humano]... Pero ahora, desconocemos el objetivo. ¿Qué quiere decir entender la mente humana? ¿Cuándo se satisfará esa pregunta? Esto es mucho, mucho más ambicioso».[10]

En 2019, al cumplirse el décimo aniversario de la ya tristemente célebre charla TED en la que Markram proclamó que un superordenador iba a cartografiar nuestro cerebro —la afirmación definitiva de la mitología sobre la IA—, tanto Scientific American (que no es precisamente enemiga de las ideas futuristas sobre la ciencia) como The Atlantic publicaron sendas investigaciones sobre lo que había salido mal.[11] En palabras de un científico: «Tenemos unos cerebros metidos en cráneos. Ahora los tenemos metidos en ordenadores. ¿Qué hemos aprendido?».[12]

Todas las preguntas tuvieron que ver con la muerte de la teoría. Y no es de extrañar. Los entusiastas del Cerebro de datos, como Markram, parecen pensar que el big data y los sistemas de aprendizaje automático que los analizan lograrán de algún modo proporcionar respuestas a las preguntas que nos hacemos sobre nosotros mismos, sobre la percepción y la inteligencia humanas que ponen en marcha esos sistemas.

Esa fe no es una ciencia novedosa, sino que es ciencia de la mala, carente de un entorno rico que facilite futuros descubrimientos.

Capítulo 18

¿El fin de la ciencia?

Aunque cada vez son más los científicos que muestran su descontento con las soluciones del cerebro de datos para las preocupaciones teóricas en curso, el ethos de la IA basado en big data ha arraigado con firmeza en la ciencia y cultura generales. Irónicamente, mientras se supone que la inteligencia general está surgiendo de la IA y sus aplicaciones hacia la investigación científica, también se le está restando importancia de manera muy visible en los roles de los científicos. El millonario, emprendedor tecnológico e inversor Peter Thiel señaló hace poco que las innovaciones parecen estar agotándose, no acelerándose.[1] En su momento, las empresas tecnológicas emergentes soñaban con la siguiente gran idea que atrajera a los inversores al Valle, pero ahora solo tienen estrategias de salida que casi siempre apuntan a su compra por parte de una gran compañía tecnológica como Google o Facebook, que de todos modos tienen bloqueada la innovación porque la IA basada en big data siempre le funciona mejor a quien posee la mayor cantidad de datos. Todo está amañado.

La cuestión es si, tal y como dice Thiel, se está produciendo un «trastorno de la cultura» o si todas las buenas ideas ya están cogidas.[2]

LA CIENCIA DE LA PASTA GANSA

Norbert Wiener, polímata, científico informático y fundador de la cibernética, nos advirtió acerca de lo que él denominó la ciencia «de la pasta gansa» en un manuscrito inédito, Invention: The Care and Feeding of Ideas [«Invención: El cuidado y la alimentación de las ideas»], que se encontró entre sus papeles después de su muerte, ocurrida en 1964.[3] A principios de los años cincuenta,Turing había completado su giro fundamental (y resultó que definitivo) hacia la idea de que la IA de nivel humano era el futuro de la

invención. Durante el mismo período, Wiener comenzó a contemplar seriamente un futuro despojado de las ideas necesarias para la IA y otros campos. La ciencia de la pasta gansa surgió con rapidez después de las dos guerras mundiales, cuando el Proyecto Manhattan había producido la bomba atómica y se habían puesto en marcha proyectos capitales y bien financiados en los ámbitos de la computación, la teoría de la comunicación y las infraestructuras. Por ejemplo, los de Bell Labs e IBM, además de otros con importantes contratistas de defensa como Raytheon. La ciencia moderna disfrutó de una historia sin precedentes de invenciones notables y en gran medida impredecibles; no obstante, a mediados de siglo, la innovación científica se había burocratizado y estaba controlada por grandes fuentes de financiación como el departamento norteamericano de defensa y los principales intereses corporativos. A Wiener le preocupaba que, en el momento mismo de su triunfo (momento también de necesidad), la cultura occidental estuviera virando hacia proyectos derivativos que de últimas amenazaban aquella próspera cultura de ideas.

Su manuscrito de principios de los cincuenta (se publicó en 1993) parece ahora profético en su lamentación: «Considero que los líderes de las tendencias actuales, que pasan de la investigación individual a una investigación controlada e industrial, están dominados, o al menos seriamente afectados, por la desconfianza hacia el individuo, lo cual a menudo implica una desconfianza en lo humano».[4]

Según el diagnóstico de Wiener, la ciencia de la pasta gansa no solo resultaba poco óptima para la cultura de la invención, sino que avanzaba directamente, y de hecho con gran felicidad, hacia lo que definió como una tendencia «antihumana». Ese parecer ha encontrado un eco en nuestra época con las críticas a la IA de gente como Jaron Lanier, que se preocupan por la erosión de la idea de humanidad inspirada por la tecnología. Las mentes colmena y los enjambres científicos harían poco por apaciguar la ansiedad de Wiener acerca de la dirección que ha tomado la ciencia. Tal y como él mismo dijo: «El efecto estadístico de una política antiintelectual llevaría a alentar la existencia de menos intelectuales y menos ideas».[5] Esas políticas antiintelectuales resultan tan evidentes en el tratamiento contemporáneo de la ciencia, basado en datos, que resulta imposible ignorar la amenaza.

Wiener señaló lo que todos sabemos o deberíamos saber; que las ideas surgen en aquellas culturas que otorgan valor a los intelectos individuales: «Las ideas nuevas se conciben en el intelecto de los científicos individuales, y en especial

tienden a originarse allí donde hay numerosos intelectos bien entrenados, y por encima de todo allí donde se valora el intelecto».[6]

En efecto, haría falta estar enajenado culturalmente para no reconocer la sabiduría de los comentarios de Wiener, que no deberían comportar la menor controversia. Igual que en la década de 1950, se siguen dedicando palabras bonitas a las innovaciones brillantes, pero la realidad es muy diferente. La cultura, tal y como temía Wiener, se ha vuelto de una forma entusiasta antiintelectual e incluso antihumana.

Esa relación con el mito es inevitable, ya que la mitología acerca del advenimiento de unas máquinas superinteligentes que vayan a reemplazar a los seres humanos hace que la ansiedad sobre el sesgo antiintelectual y antihumano resulte irrelevante. La cuestión principal del mito es que el antihumanismo representa el futuro; está integrado en el avance de la tecnología existente.

Cuesta imaginar un meme cultural que resulte tan corrosivo para la prosperidad futura de una manera tan directa, y paradójicamente tan hostil, también de manera directa, para la invención o el descubrimiento mismos de una teoría factible sobre la inteligencia general. Que esa teoría vaya a aparecer en la investigación y desarrollo futuros es un enigma, pero lo que sí se puede reconocer es la amenaza de una cultura de las ideas cada vez más anémica y que militará en contra de cualquier descubrimiento de ese tipo. El efecto general del mito en ese contexto consiste solo en promover la IA y, en efecto, en llevar la investigación científica misma hacia un modo tecnocéntrico, donde se desalentará y se dejará de reconocer de manera sistemática la invención genuina —si es que esta llega a darse, tratándose de algo raro a lo largo de los siglos y aún más en la actualidad.

UNA APUESTA POR LAS IDEAS

Wiener señaló que la economía del beneficio corporativo dificulta la inversión en una cultura de las ideas genuina, ya que las apuestas iniciales sobre una idea son esencialmente malas y su valor total solo se torna aparente con el paso del tiempo.

Por decirlo de manera simple, las ideas novedosas no se pueden predecir, de modo que representan un compromiso económico e intelectual hacia una cultura de la prosperidad que no garantiza ganancias a corto plazo. En otras palabras, cabe anticipar que la consolidación de la web como gran tecnología también tienda a inclinar la labor de la IA hacia aplicaciones débiles que impulsen una curva de beneficios, mientras que los inventos (aún desconocidos) recibirán poca atención.

Como prueba de esta afirmación, pensemos en la escasa inversión que se ha dedicado a explorar nuevos caminos hacia la inteligencia artificial general en oposición a, por ejemplo, las aplicaciones de aprendizaje profundo para juegos. Estas últimas representan con claridad un callejón sin salida para la inteligencia artificial general, tal y como están comenzando a admitir hasta los investigadores de aprendizaje profundo —recelosos, sin duda, de que la IA viva otro invierno infame de la mano de una nueva burbuja—. La cultura consiste en exprimir los beneficios de las frutas que cuelgan al alcance de la mano mientras se continúa dando vueltas a la mitología de la IA, una estrategia que garantiza el desencanto si no llega una innovación conceptual radical.

Wiener observó con ironía que el mundo satírico creado por Swift en Laputa, donde hay una máquina que hace evolucionar la ciencia de manera «automática», tuvo una cierta presencia intelectual en la ciencia de la pasta gansa de los años cincuenta, con el resultado inevitable de que alienó todavía más la cultura de la invención. Wiener se mostró preocupado en especial ante algunas versiones tempranas de algo que ahora es parte integrante de la mitología de la IA, el hecho de que la mente humana esté siendo reemplazada por los programas informáticos: «El deseo contemporáneo por encontrarle un reemplazo mecánico a la mente humana tiene unos límites bien definidos. Allí donde las tareas realizadas por un individuo se puedan comprender con atención y nitidez, no costará encontrar un sustituto razonablemente adecuado o bien un dispositivo mecánico, o una organización en la que las mentes humanas se unan como si fueran partes de ese dispositivo».[7]

Por supuesto, la observación de Wiener es una reafirmación perfecta de la mitología de la IA y su nocivo efecto sobre la humanidad, con mentes colmena en la red y enjambres de científicos en la investigación. Se nos podría perdonar que no nos «quedemos esperando» invenciones y descubrimientos cuando disponemos de los superordenadores Blue Gene de IBM para jugar, pero lo que resulta imperdonable es el intento deliberado por reducir a la persona, tal y como

dice Lanier —denigrando y arrebatándole su valor e importancia a la mente humana misma—. Esa estrategia resulta extraordinariamente contraproducente y estúpida.

A continuación, Wiener relacionó su crítica con las fantasías mecánicas populares, a menudo objeto de burla de aquellos escépticos que siempre se muestran recelosos ante el sueño de las máquinas. (Ya hemos visto la sátira de la ciencia mecánica a cargo de Jonathan Swift, cuando tratamos la exposición de Peirce sobre el desarrollo inicial del razonamiento automatizado.) Wiener prosiguió: «Sin embargo, el uso de la mente humana para desarrollar ideas nuevas de verdad es un fenómeno novedoso en cada ocasión. La expectativa de obtener ideas nuevas con un contenido real a través de multiplicar las actividades humanas de baja cualificación, y a través de una remodelación azarosa de ideas ya existentes, sin el liderazgo de una mente de primera categoría, que seleccione esas ideas, es una variante de la falacia de los monos y la máquina de escribir que ya aparece bajo una manifestación ligeramente más simple en "Un viaje a Laputa", de Swift».[8]

Aquí, Wiener captura a la perfección la fantasía por la que Henry Markram ha invertido mil millones de euros en la mitología de la IA, construyendo un cerebro que use redes neuronales y superordenadores (y las teorías neurocientíficas existentes). Ojalá alguien hubiera expuesto y evitado esas ideas. De hecho, el giro contemporáneo hacia la IA parece haber otorgado a esas fantasías un papel aún más central en nuestra cultura, con el éxito de unas aplicaciones predeciblemente débiles pero brillantes, que se publicitan como un anticipo de ese futuro que (ay) estará dominado por las máquinas superinteligentes. Esa visión de la inteligencia artificial general es pura fachada y mitología. Lo más probable es que no haya nadie capaz de comprender con claridad los problemas fundamentales, y mucho menos que vaya a encontrar las ideas necesarias para que se dé un progreso verdadero. Esta comparación podría provocar sonrisas de suficiencia, pero no obstante resulta adecuada: es un hermoso mundo nuevo.*i Hay que reconocerle a Wiener que se lo vio venir.

CADA VEZ MÁS DÉBIL

Situando la IA como marco de inferencia, estamos siendo testigos, en efecto, de la evolución de una subespecie en la IA inductiva que puede desempeñarse bien en los entornos restrictivos y basados en datos, pero que de manera necesaria carece de la capacidad para aprender a tener sentido común y adquirir una comprensión genuina. Es simplemente increíble que estemos ligando el futuro de la mente humana —no tan limitada— a un mayor desarrollo de ese tipo de IA.

No se trata solo de que ese enfoque carezca por completo de la inteligencia general necesaria para realizar algún progreso intelectual y real en la cultura moderna, sino que, puesto que la inducción es diferente a la abducción de manera comprobable, ya sabemos que no existe ningún puente entre la una y la otra.

Todas las proclamas de Ray Kurzweil acerca de ese progreso inevitable no pueden deshacer la verdad una vez que la conocemos. Deberíamos ser sinceros, porque reconocer la verdad formará parte integral del plan de acción para salir adelante.

Resumiendo, para comenzar no hay manera de que la IA actual «desarrolle» una inteligencia general en ausencia de algún descubrimiento fundamental. Limitarse a decir que «estamos llegando» nos conduce a la bancarrota científica y conceptual, y atiza las llamas de unas fuerzas antihumanas y antiintelectuales interesadas (al parecer) en controlar y predecir unos resultados que, entre otros motivos, intentan maximizar el beneficio a corto plazo dirigiendo el debate hacia el carácter inevitable de la cuestión. Los sujetos listos alteran el curso de las cosas; una de las maneras de hacer que el futuro resulte más predecible consiste tan solo en denigrar y eliminar cualquier valor que hayamos depositado en la inteligencia individual.

CIRCULE: AQUÍ NO HAY NADA QUE VER

La sugerencia de que nos hemos adentrado en un callejón sin salida cultural podría sonar ficticia y fantástica si no hubiera tantos proveedores de la mitología de la IA defecándose felices ante los micrófonos en la preocupación de Wiener por «cuidar y alimentar las ideas» a la vez que ensalzan la inevitabilidad de la

IA. Mientras los científicos de la IA y mitólogos a tiempo parcial como Stuart Russell continúan advirtiéndonos de que no descartemos el ingenio humano en la búsqueda de una teoría futura de la inteligencia artificial general, son muy escasos los líderes de la cultura contemporánea que están promoviendo una agenda en la que pueda prosperar el ingenio humano.

Dados los objetivos (o miedos) manifiestos de crear en efecto a un nuevo superser, eso resulta asombroso. Sin duda, hoy en día nos irían bien un Einstein o dos. (Uno se pregunta qué tal le iría a Turing en estos momentos.)

De nuevo, el dogma de la mitología de la IA es el lugar donde esto se manifiesta de manera más evidente. En cualquier cálculo sobre el futuro de la inteligencia artificial general, la responsabilidad recae de lleno en los mitólogos de la IA que presagian el advenimiento de la IA de nivel humano para explicar nuestras acciones de cara a que las cosas avancen.

Quizá podríamos comenzar por reconocer con franqueza que el aprendizaje profundo es un callejón sin salida, pues pertenece a una IA basada en datos, y tanto da la cantidad de dólares de publicidad que pueda llevar a las arcas de las grandes empresas tecnológicas. También podríamos dar más voz a una realidad que cada vez más científicos de IA están reconociendo, si bien a regañadientes: que, igual que en períodos previos de gran excitación a vueltas con la IA, nadie tiene la menor idea sobre cómo construir una inteligencia artificial general.

El sueño continúa siendo mitológico precisamente porque la ciencia contemporánea no ha llegado a entenderlo en absoluto. ¿En qué otro lugar podemos librarnos de ese mito, más que en la misma ciencia de la IA?

JOHN HORGAN Y LA INQUIETANTE SUGERENCIA DEL FINAL

El fantasma de una sociedad puramente tecnocrática en la que la ciencia, que en su día nos proveyó de descubrimientos e invenciones radicales y revolucionarios, pase a desempeñar el papel del técnico de laboratorio que se dedica a ajustar las ruedecillas de los «cerebros gigantes» de los

superordenadores fue sugerido en un primer momento por John Horgan, redactor de Scientific American. A mediados de la década de 1990, en su popularísima obra El final de la ciencia, Horgan se preguntó si el agotamiento aparente de la investigación básica en el mundo de la ciencia era inevitable, por el simple hecho de que la época de los principales descubrimientos había quedado atrás.[9]

Es la mitad de la pregunta que Thiel se hace en la actualidad: ¿la cultura se ha trastornado, ha tomado un rumbo que asfixia las nuevas ideas, tal y como temía Wiener, o nos hemos quedado sin ideas básicas porque ya las hemos encontrado todas? Esta última posibilidad representaría «El fin» en un sentido esencial —así que podemos rezar porque la cultura se haya limitado a abrazar una respuesta tecnológica universal para las preguntas básicas y porque la asfixia de la inteligencia humana tan solo sea un daño colateral—. Existe al menos una manera hipotética de reparar una cultura científica trastornada; escapar de ese mundo de Tron en el que se han acabado las ideas representa una nueva pesadilla.

La pregunta de Thiel resulta fundamental no solo para el futuro de la IA, sino también para el de la humanidad, y por desgracia hay pruebas que sustentan ambas hipótesis. Por un lado, la promoción alegre del mito y de su primo en la ciencia de enjambre —igual que la celebración de las mentes colmena antes de esta— parece sugerir que la sociedad moderna en efecto se ha adentrado silbando en una especie de alienación respecto a los valores fundamentales, tal y como Wiener presagió.

Por otro lado, la pregunta sobre si no tenemos otra elección, tal y como asegura Horgan, presenta la inquietante posibilidad de que en este momento, más de trescientos años después de la revolución científica, hayamos recogido toda la fruta que las teorías física e informática dejaron al alcance de nuestras manos. Según esta visión, ya hemos descubierto todo lo que se podía descubrir sobre la física, primero con las leyes de Newton y más tarde con la relatividad de Einstein y el desarrollo de la mecánica cuántica durante el siglo XX. Los avances que le queden a la física consistirán en gran medida en rellenar huecos y añadir detalles a la teoría ya existente, y sin duda se pondrán a prueba las predicciones de esas teorías con tecnologías cada vez más caras y voluminosas, como los supercolisionadores. Bienvenidos a Maquinalandia.

Cualquiera de esas posibilidades negativas respaldaría la sugerencia de Markram según la cual Einstein hoy sería indeseable, y no tendría nada que hacer (salvo

quizá contribuir a la ciencia de datos). La inevitabilidad del advenimiento de la superinteligencia queda aquí colgando boca abajo, porque los seres humanos, tan notables a la hora de descubrir los bloques de construcción fundamentales del universo, deben ahora retroceder unos pasos y limitarse a observar mientras la cultura los hace pasar de descubridores a técnicos. Ocuparse de los ordenadores es la versión contemporánea del Voltaire que cultivaba su jardín. El trabajo serio ha llegado a su fin. Los seres humanos no deberían haber sido tan listos.

Horgan también sugiere que algunos sueños, como el de la explicación científica completa para la conciencia humana, podrían haber sido, en todo caso, demasiado complejos, y tremendamente remotos. En ese caso, tendremos el resultado infeliz de evidenciar que la informática —y sus teorías actuales— trepa inexorable en el camino de la ciencia y de todo lo demás, mientras que se pondrá discretamente a dormir a los sueños prometeicos de una neurociencia completa, que como mucho aparecerá en futuros ficticios como el que mostraba Ex Machina.

Sin duda existe la posibilidad de que hayamos dejado atrás los avances científicos capitales, en cuyo caso deberíamos esperar un tratamiento técnico y superficial de los temas centrales usando teorías ya existentes, mientras la mitología de la IA se convierte en el nuevo foco de nuestros propósitos para el futuro, uno nihilista y falso. Tal y como también ha sugerido Lanier, podemos hacer realidad ese futuro si nos limitamos a cincelar la inteligencia y la singularidad humanas hasta encorvarnos lo suficiente como para adecuarnos a ese futuro presidido por los ordenadores.

No es que Horgan se mostrara encantado con su propia e inquietante sugerencia, pero da la sensación de que, desde los noventa, la informática aplicada le viene otorgando cada vez mayor credibilidad —si no en la realidad, al menos en la práctica observada.

En uno u otro caso, deberíamos tomarnos con seriedad el hecho de que estamos siguiendo el camino equivocado, en gran parte porque intentamos activamente ocultar un déficit clave, la falta de una cultura humana próspera, a base de retórica sobre el inevitable ascenso de las máquinas. Ni a Eugene Goostman se le hubiera ocurrido un camino mejor para no obtener resultados.

NUESTRA ELECCIÓN

Si la lectura que realizó Horgan es cierta y nuestro futuro ha llegado a su fin, resulta inevitable que nos dejemos llevar a la deriva hacia el detalle técnico. Sin embargo, el desvarío de la cultura, extendido en buena parte por el mito (y el ascenso de la informática ubicua), mantiene viva la posibilidad de que, si nos liberamos de los mitos de la tecnología moderna, estimulemos el progreso reinvirtiendo en la perspicacia, la innovación y las ideas humanas.

Es evidente que estoy a favor de esta última interpretación. Y soy optimista, en gran medida porque, tal y como hemos visto, desde una posición puramente científica, tenemos todos los motivos para rechazar la marcha inevitable y lineal hacia la inteligencia artificial general (y más allá).

Para desatar ese nudo gordiano debemos comenzar por librarnos del mito en su apariencia actual, responsable de infectar la cultura de forma tan dominante que hace falta un largo debate sobre la necesidad de una nueva teoría neurocientífica que reconduzca nuestros esfuerzos, una cuestión que ya debería haber quedado clara y que no necesita discusión.

LA CONFIANZA COMO RECONOCIMIENTO DE LOS LÍMITES DE LOS SISTEMAS INDUCTIVOS

Irónicamente, los límites de la IA contemporánea se encuentran implícitos en los debates actuales sobre la automatización y la confianza. Hay una tendencia de moda, entre los pensadores de la IA, que se preocupa por la denominada «IA amigable», los sistemas de confianza y demás cuestiones éticas como el sesgo problemático. En otras palabras, los sistemas que no entienden, pero continúan actuando, se han convertido en un motivo de preocupación.

Esto se proyecta sobre el mito en un ángulo incómodo: que las cuestiones sobre el valor humano estén infectando lo que en su día fueron asuntos puramente científicos se debe a que los sistemas, pese a ser idiotas, no han dejado de abrirse

camino por las utilidades empresariales, de consumo y gubernamentales.

Los coches sin conductor son un buen ejemplo de ello. Está muy bien hablar sobre los avances en el reconocimiento visual de los objetos hasta que, en algún punto de la larga cola de consecuencias imprevistas y, por consiguiente, no incluidas en los datos de entrenamiento, tu vehículo embiste un autobús lleno de gente en su intento por esquivar una torre eléctrica (es algo que ha pasado). Fijémonos también en los problemas con el sesgo y el reconocimiento de imágenes: Google Photos estampó la etiqueta de GORILA sobre la foto de dos afroamericanos. Tras esa bomba de neutrones de los desastres de relaciones públicas, Google solventó el problema... deshaciéndose de las imágenes con gorilas en el conjunto de entrenamiento que usaba su sistema de aprendizaje profundo.

Así, los límites de esa IA inductiva que carece de una comprensión genuina van formando parte creciente del debate sobre la IA porque estamos precipitándonos al hacer que las máquinas entren en servicio, pese a su falta de entendimiento, en áreas importantes de la vida humana. Eso también es consecuencia de la mitología de la IA, que muestra una inclinación recurrente a no esperar la llegada de ideas o descubrimientos legítimos, siempre demasiado ansiosa por aumentar el predominio de las tecnologías de la IA en todas las áreas posibles de la vida.

Irónicamente, esta tendencia preocupante podría ayudar a estimular una mejor comprensión de los límites fundamentales de la IA —o, en el peor de los casos, de sus límites actuales e inevitables.

En otras palabras, mucha tinta se está vertiendo en nombre del mito para describir lo que vendría a ser la estupidez de las máquinas. Nadie parece reparar en que el resultado es una consecuencia necesaria y predecible de que los sistemas inductivos se hayan hecho pasar por un camino hacia la inteligencia.

Russell apunta hacia el problema de «alineación», cuestión que de repente ha cobrado una importancia fundamental en la IA y que consiste en armonizar los sistemas actuales y futuros de la IA con nuestros propios propósitos e intereses. Pero ese problema no surge, tal y como sugiere Russell, porque los sistemas de IA se hayan vuelto tan inteligentes en tan poco tiempo, sino porque nos hemos dado demasiada prisa en otorgarles posiciones de autoridad en diversas áreas de la sociedad humana, y sus limitaciones inherentes —que siempre estuvieron ahí

— han cobrado importancia.

Tengo la esperanza de que el alejamiento actual de la Singularidad, en pos de la preocupación práctica de conceder una autoridad real a la IA —a, reconozcámoslo, unas máquinas carentes de mente—, acabará provocando que la inteligencia y el valor humanos sean objeto de un aprecio renovado.

Prestar atención al problema de alineación podría llevar a que se reflexione más sobre el acrecentamiento —la manera en que podemos usar de la mejor manera a unos eruditos idiotas cada vez más potentes para promover nuestros propios objetivos, incluida la búsqueda del progreso científico.

PARA CONCLUIR

El marco de pensamiento inferencial que he presentado en este libro clarifica el proyecto de expansión de la inteligencia artificial actual hacia la inteligencia artificial general: debe tender un puente a un tipo de inferencia diferente, que en la actualidad no se puede programar. También sirve como una guía para explorar los límites entre la mente y la máquina, y para facilitar que la interacción entre personas y máquinas, que, por supuesto, está aquí para quedarse, sea mejor y más segura. Sigue siendo verdad que la tecnología a menudo actúa como una prótesis de las capacidades humanas, como sucede con el microscopio y el telescopio. La IA ha de desempeñar al menos ese papel, pero la mitología acerca de la llegada de la superinteligencia debería mantenerse en el apartado de enigmas científicos. Si deseamos perseguir un misterio científico de manera directa, debemos, en todo caso, invertir en una cultura que incentive las ideas intelectuales —las vamos a necesitar en caso de que exista algún camino posible a la inteligencia artificial general.

Tal y como Frankenstein fue, en realidad, una exploración del aislamiento espiritual (un problema que afectó profundamente a Mary Shelley y a su marido, Percy Shelley), las cuestiones de mayor calado que se encarnan en el mito de la IA no son de naturaleza técnica ni científica, sino que tienen en cuenta nuestro esfuerzo en curso por encontrar un sentido y por forjarnos caminos de futuro en un mundo en cambio permanente. No ganamos nada entregándonos al mito, que

no puede ofrecer soluciones para nuestra naturaleza más que en el sentido manifiestamente negativo de menospreciar el potencial de la humanidad y limitar sus posibilidades futuras.

El problema de la inferencia, igual que el problema de la conciencia, hunde sus raíces en el núcleo de una serie de misterios capitales que siguen teniendo validez, y esta arraigado en nuestra comprensión de todas las demás cosas. No debería sorprendernos que las partes de la mente que aún no hemos descubierto se resistan a una respuesta tecnológica. Es posible que, tal y como temía Horgan, nos hayamos quedado sin ideas. En tal caso, el mito representa el alejamiento final e irrecuperable respecto a la posibilidad humana —un cuento de hadas oscuro y reconfortante, la pretensión de que algo diferente, algo vivo y magnífico, deberá sin duda elevarse de nuestras cenizas.

Y, si no nos hemos quedado sin ideas, tenemos que llevar a cabo la tarea dura y deliberada de volver a invertir en una cultura de la invención y de la prosperidad humana. Porque necesitaremos de nuestra propia inteligencia general para encontrar caminos hacia el futuro, y para que ese futuro sea mejor que el pasado.

AGRADECIMIENTOS

Han sido muchas las personas involucradas, formal o informalmente, en la creación de este libro. Ante todo, en el equipo de Harvard University Press, me gustaría dar las gracias a Jeff Dean, ahora en Hackett Publishing, por el apoyo inicial que le dio al proyecto cuando era editor de HUP. Con amabilidad, Jeff me hizo volver atrás sobre mis ideas cuando comenzaba a ponerlas por escrito, y le estoy agradecido por su paciencia a la hora de sortear los primeros baches. James Brandt, Joy de Menil y Julia Kirby tomaron su testigo y mejoraron el texto cada uno a su manera. Graciela Galup aportó un diseño elegante y Colleen Lanick y su equipo de prensa encontraron maneras creativas de promocionarlo.

Me gustaría dar reconocimiento también a otras personas que ayudaron a que el libro viera la luz. Pasé parte del período de redacción en el extranjero, y algunos sucesos en Estados Unidos me pusieron las cosas difíciles en ciertos momentos. Mi agradecimiento a Bernard Fickser por proporcionar financiación a través de un generoso contrato en la empresa emergente que creamos juntos, así como por su apoyo en materias demasiado diversas como para que las enumere aquí. Todd Hughes leyó una versión inicial del manuscrito e hizo varias sugerencias provechosas. Mi agente, Roger Freet, ahora en la agencia literaria Folio, se mostró incansable aspectos diversos sobre la escritura y la publicación, y agradezco mucho sus propuestas.

Asimismo me gustaría dar las gracias a Anna Samsonova por aguantarme durante aquellos largos días en Europa, y por ayudarme con el embrollo interminable de aprender ruso mientras intentaba formular mis ideas en inglés. También me siento agradecido respecto a mis amigos escritores de Seattle (ya sabéis quiénes sois), que realizaron contribuciones muy valiosas, aunque, de algún modo, todavía no lo saben.

Por último, a mi amigo John Horgan, el gran escritor de ciencia que ha inspirado a tantos, muchas gracias por creer desde muy pronto en este libro, y por tu apoyo continuado.

Sobre el autor

ERIK J. LARSON

Erik J. Larson es un científico experto en computación y emprendedor tecnológico. Fundador de dos nuevas empresas de IA financiadas por DARPA (la Agencia de Proyectos de Investigación Avanzados de Defensa de EE.UU.), actualmente trabaja en cuestiones fundamentales acerca del procesamiento del lenguaje natural y el aprendizaje automático. A través de sus investigaciones en el think tank IC2 de la Universidad de Texas ha experimentado con los límites técnicos de la IA y, en su faceta de escritor, ha publicado artículos en The Atlantic y otros medios profesionales.

NOTAS

Introducción

1. No pretendo sugerir que los investigadores no se hayan peleado con la abducción en la IA: lo han hecho. Durante las décadas de 1980 y 1990, los investigadores trabajaron en algunas aproximaciones lógicas a la abducción, llamadas programación lógica abductiva (ALP en sus siglas inglesas), pero esos sistemas eran abductivos solo nominalmente, ya que se basaban en la deducción y no en una verdadera abducción. No tuvieron éxito, y se abandonaron con rapidez en cuanto el trabajo de la IA progresó con la llegada de la web. En tiempos más recientes, más o menos desde 2010 y hasta el presente, se han adoptado varios enfoques probabilísticos (en especial, bayesianos) como posibles rutas hacia una inferencia abductiva genuina. No obstante, esos sistemas tampoco son tratamientos completos de la abducción. En vez de ser enfoques deductivos disfrazados, como sus predecesores, son enfoques inductivos o probabilísticos disfrazados. Cuando hablo de abducción no me refiero solo al nombre, y los sistemas que la usen pero que no resuelvan el problema no nos ayudarán a progresar con la IA. Es algo que explicaré en las páginas que siguen.

Capítulo 1: El error de la inteligencia

1. A. M. Turing, «Computing Machinery and Intelligence», Mind 59, n.º 236 (octubre de 1950), pp. 433–460.

2. A. M. Turing, «On Computable Numbers, with an Application to the Entscheidungsproblem», Proceedings of the London Mathematical Society, vols. 2–42, issue 1 (enero de 1937), pp. 230–265.

3. A. M. Turing, Systems of Logic Based on Ordinals, tesis doctoral,

Universidad de Princeton, 1938, p., 57.

4. Gödel también demostró que el añadido de unas reglas serviría para reparar la incompletitud de algunos sistemas, pero que el nuevo sistema, con sus reglas adicionales, tendría otros puntos ciegos, y así una y otra vez. En ello se centró precisamente Turing en sus últimos trabajos sobre los sistemas formales y la completitud.

5. Para conocer los resultados originales de la incompletitud, véase Kurt Gödel, «Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I», Monatshefte für Mathematik Physik 38 (1931), pp. 173–198. Existe una traducción al inglés en Kurt Gödel, Collected Works, vol. 1: 1929–1936, eds. Kurt Gèodel, Kurt Gödel y Solomon Feferman, Oxford, Oxford University Press, 1986.

6. Aquí voy a usar los términos «formal», «matemático» y «computacional» de modo que sean intercambiables entre sí. La terminología no es imprecisa, aunque técnicamente todos los sistemas matemáticos o computacionales se conozcan como sistemas formales. Confío en que esto no resulte confuso pero, en cualquier caso, tanto el término «matemático» como el término «computacional» se refieren a los sistemas formales, que disponen de un vocabulario de símbolos bien definido y de reglas para manipular esos símbolos. Eso comprende la aritmética, así como los lenguajes informáticos, y su carácter por completo general se adecua a los propósitos del debate.

Capítulo 2: Turing en Bletchley

1. Los trabajos iniciales de Turing, Good y Shannon sobre el ajedrez computerizado explotaron una técnica conocida como «minimax», que realizaba sus movimientos intentando minimizar las pérdidas del jugador y maximizando a su vez su ganancia potencial. Esa técnica desempeñó un papel prominente en versiones futuras del ajedrez de ordenador, y sigue representando el punto de partida para los sistemas de ajedrez informático mucho más potentes que se utilizan en la actualidad.

2. Todos los lenguajes de programación completos según los conocemos hoy en día, como C++ y Java, utilizan estas operaciones básicas, aparecidas en los inicios de la computación, aunque ideas como la de la programación orientada a objetos y demás métodos para estructurar el software aparecieron más tarde, con la ciencia informática. Aun así, las estructuras básicas de control de todo el código informático llegaron muy pronto, con las primeras máquinas electrónicas completas. La percepción sobre cómo estructurar y controlar las máquinas a través de programas es sin duda responsable del éxito inmediato de esos sistemas en su aplicación a problemas iniciales como el ajedrez.

3. Bajo la dirección de la Escuela Gubernamental de Códigos y Cifrados (GC and CS, en sus siglas inglesas).

4. Hugh Alexander, por ejemplo, fue un campeón nacional de ajedrez que ayudó en los afanes de Bletchley.

5. Los alemanes le añadieron rotores que generaban largas secuencias de texto libre cifrado, junto con unas instrucciones iniciales más complicadas (y también cifradas) para descifrar las comunicaciones.

6. Véase la excelente biografía que Andrew Hodges dedicó a Turing para acceder a un examen en profundidad del papel que Bletchley Park y Turing desempeñaron en la segunda guerra mundial. Andrew Hodges, Alan Turing: The Enigma, Nueva York, Vintage, 1992.

7. Joseph Brent, Charles Sanders Peirce: A Life, Bloomington, IN, Indiana University Press, 1993, p. 72.

8. Hodges, Enigma, p. 477.

9. François Chollet, «The Implausibility of Intelligence Explosion», Medium, 27 de noviembre de 2017.

10. Para un tratamiento matemático del teorema «No Free Lunch», véase David Wolpert y William G. Macready, «No Free Lunch Theorems for Optimization»,IEEE Transactions on Evolutionary Computation 1, n.º 1 (1997), pp. 67–82.

11. De hecho, la expresión «inteligencia artificial» fue acuñado en 1955 por John McCarthy, científico informático de Stanford, uno de los pioneros de la IA y

miembro de la Conferencia de Dartmouth, que lanzó la disciplina de manera oficial un año después, en 1956.

Capítulo 3: El error de la superinteligencia

1. Irving John Good, «Speculations Concerning the First Ultraintelligent Machine», Advances in Computers 6 (1965), pp. 31–88.

2. Nick Bostrom, Superintelligence: Paths, Dangers, Strategies, Oxford, Oxford University Press, reed. 2017, p. 259.

3. John Von Neumann, Theory of Self-Reproducing Automata, ed. Arthur W. Banks, Urbana, University of Illinois Press, 1966, quinta conferencia, p. 78.

4. Daniel Kahneman, Pensar rápido, pensar despacio, Barcelona, Debate, 2012.

5. Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Nueva York, Viking, 2019, p. 37.

6. Kevin Kelly, What Technology Wants, Nueva York, Penguin, 2010.

7. Resulta extraño, o quizá refrescante, que desde ese momento Kelly se haya distanciado del mito de la IA. En un artículo para Wired de 2017 sostiene que, al fin y al cabo, la «inteligentización» no ha conducido a la superinteligencia. Señala que la inteligencia es variopinta y polimorfa, y que animales en apariencia no inteligentes, como las ardillas, recuerdan la localización de las —posiblemente— miles de nueces que han enterrado para consumir más adelante, hito que las personas sin duda no podrían imitar. El título de ese artículo, «The AI Cargo Cult: The Myth of Superhuman AI» («El culto cargo de la IA: El mito de una IA sobrehumana») resulta revelador. Nuestra incapacidad para fijar una definición de inteligencia con la que trabajar (y no hablemos ya de la de superinteligencia) podría sugerir que el punto final que los futurólogos ven como resultado inevitable de la IA resulta de hecho confuso, y que se trata de otra simplificación que deja un amplio margen para la mitología y la especulación.

8. Russell, Human Compatible, pp. 7–8.

Capítulo 4: La singularidad, ayer y hoy

1. Murray Shanahan, The Technological Singularity, Cambridge, MA, MIT Press, 2015, p. 233.

2. Como ya hemos visto, a finales de los años treinta la ciencia de las máquinas universales de Turing ya estaba bien establecida. Los ordenadores como aparatos electrónicos aparecieron más tarde, aparejados con el desarrollo de la tecnología de la comunicación y la aparición de los interruptores de relé y demás.

3. Técnicamente, Vinge introdujo el término de singularidad tres años antes, en un artículo para la revista Omni de enero de 1983 que llevó por título «First Word» («Primera palabra»). No obstante, es común rastrear la palabra y el uso que Vinge le dio hasta su novela de ciencia ficción Naufragio en el tiempo real (Barcelona, Nova, 1989), en cuyo argumento se da un desarrollo completo del concepto.

4. Vernor Vinge, «The Coming Technological Singularity: How to Survive in the Post-Human Era», Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace, ed. G. A. Landis, NASA Publication CP-10129, 1993, pp. 11–22.

5. Ray Kurzweil, La singularidad está cerca. Cuando los humanos trascendemos la biología, Berlín, Lola Books, 2021.

6. Ray Kurzweil, «The Singularity: A Talk with Ray Kurzweil», entrevista con The Edge, introducción de John Brockman. 24 de marzo de 2001. https://www.edge.org/conversation/ray_kurzweil-the-singularity.

7. Hubert L Dreyfus, What Computers Still Can’t Do: A Critique of Artificial Reason, Cambridge, MA, MIT Press, 1992, p. ix.

Capítulo 5: La comprensión del lenguaje natural

1. John McCarthy, M. Minsky, N. Rochester y C. E. Shannon, «A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence», agosto de 1955.

2. Gary Marcus y Ernest Davis, Rebooting AI: Building Artificial Intelligence We Can Trust, Nueva York, Pantheon Books, 2019, p. 1.

3. Massimo Negrotti, ed., Understanding the Artificial: On the Future Shape of Artificial Intelligence, Berlin Heidelberg, Springer-Verlag, 1991, p. 37.

4. Véase John R. Pierce et al. , Language and Machines: Computers in Translation and Linguistics. Informe del Comité consultivo para el procesamiento avanzado del lenguaje, Academia Nacional de las Ciencias. Publicación 1416 del Consejo Superior de Investigaciones Científicas de EEUU. 1966.

5. Sergei Nirenburg, H. L. Somers y Yorick Wilks, eds., Readings in Machine Translation, Cambridge, MA, MIT Press, 2003, p. 75.

6. Para leer un debate ameno sobre los problemas iniciales de la traducción automática véase: John Haugeland, Artificial Intelligence. The Very Idea. 1989. Cambridge, MA, MIT Press, El comentario de Yehoshua Bar-Hillel aparece en la página 176.

7. Véase, por ejemplo, el recuento de Hubert L. Dreyfus sobre los fallos de los inicios de la traducción automática en: Hubert L Dreyfus, What Computers Still Can’t Do: A Critique of Artificial Reason, Cambridge, MA, MIT Press, 1992, p. ix.

8. Para saber más sobre DENDRAL, véase: Robert K. Lindsay, Bruce G. Buchanan, E. A. Feigenbaum y Joshua Lederberg, «DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation», Artificial Intelligence 61, n.º 2 (1993), pp. 209–261. Para saber más sobre, véase: B. G. Buchanan y E. H. Shortliffe, Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project, Reading, MA, Addison-Wesley, 1984.

9. Véase un buen debate acerca del programa ELIZA de procesamiento de

lenguaje natural en la IA en: James Allen, Natural Language Processing, San Francisco, Benjamin /Cummings Publishing Company, 1995. El diálogo aparece en la página 7.

10. Explicaré los problemas de la actuación de Goostman con mucho mayor detalle en la segunda parte del libro.

Capítulo 6: De la IA como tecnología kitsch

1. El término de «tecnociencia» es aquí un anacronismo, pero captura con exactitud y en retrospectiva las ideas que comenzaron a cobrar forma en el siglo XIX a consecuencia de la revolución científica. En realidad fue el filósofo belga Gilbert Hottois quien acuñó la palabra «tecnociencia» en los 1970.

2. Friedrich Wilhelm Nietzsche, La gaya ciencia, Madrid, Akal, reed. 2019.

3. Fiódor Dostoyevski, Memorias del subsuelo, Madrid, Akal, reed. 2018.

4. Compartieron esas opiniones los pensadores franceses que influyeron en Comte, entre los que se cuenta Marie Jean Antoine Nicolas Caritat, el marqués de Condorcet (a quien por lo general se conoce simplemente como Condorcet). Antes de la revolución de la filosofía francesa que la revolución científica trajo consigo, filósofos ingleses como Francis Bacon abrazaron ideas similares al considerar que la ciencia y el progreso eran filosóficamente esenciales.

5. Hannah Arendt, La condición humana, Barcelona, Paidós, reed. 2016.

Capítulo 7: Simplificaciones y misterios

1. B. F. Skinner,Walden dos (1948), Madrid, Ediciones Martínez Roca, reed. 1984.

2. Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Nueva York, Viking, 2019, p. 8.

3. Dan Gardner, Future Babble: Why Expert Predictions Are Next to Worthless, and You Can Do Better, Nueva York, Dutton, 2011.

4. Martin Ford, Architects of Intelligence: The Truth About AI From the People Building It, Birmingham, RU, Packt Publishing, 2018, p. 20.

5. Ray Kurzweil, La singularidad está cerca. Cuando los humanos trascendemos la biología, Berlín, Lola Books, 2021, p. 25.

6. Alasdair MacIntyre, Tras la virtud, Barcelona, Crítica, reed. 2004.

7. Véase: Michael Polanyi, Personal Knowledge. Towards a Post-Critical Philosophy (1958), Abingdon-on-Thames, RU, Routledge & Kegan Paul, ed. corregida en 1962, cap. 5.

8. Hubert Dreyfus,What Computers Still Can’t Do: A Critique of Artificial Reason, Cambridge, MA, MIT Press, 1992.

9. Gary Marcus y Ernest Davis, Rebooting AI: Building Artificial Intelligence We Can Trust., Nueva York, Pantheon Books, 2019.

10. Hector Levesque, Common Sense, the Turing Test, and the Quest for Real AI, Cambridge, MA, MIT Press, 2017.

11. Erik J. Larson, «Questioning the Hype About Artificial Intelligence», The Atlantic, 14 de mayo de 2015.

12. Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Nueva York, Viking, 2019, p. 9.

13. Ibid. , p. 41.

14. Ibid. , pp. 16-17.

15. Ford, Architects, pp. 232, 234.

16. La escena en la que Ava huye de las instalaciones de Nathan y levanta la

mirada hacia la luz del sol para ver en color aborda un problema que en la IA y en los círculos de la filosofía de la mente se conoce como «El cuarto de Mary». El misterio consiste en saber si una científica llamada Mary, que estudia los colores y que (hipotéticamente) conoce todos los datos científicos acerca del color (longitudes de onda que impactan sobre las neuronas del cerebro, etc.), pero que vive en un cuarto en blanco y negro, aprendería de hecho algo nuevo al ver los colores de verdad por primera vez. En otras palabras, ¿representa nuestra experiencia de ver el color un añadido respecto a los cálculos sobre la cuestión? Garland sugiere que, en efecto, Ava ha aprendido algo nuevo. Presumiblemente, eso demuestra que también posee una mente consciente.

17. Eliezer Yudkowsky, «Artificial Intelligence as a Positive and Negative Factor in Global Risk», en Global Catastrophic Risks, eds. Nick Bostrom y Milan M. Ćirković, Nueva York, Oxford University Press, pp. 308–345.

18. Jaron Lanier, You Are Not a Gadget: A Manifesto, Nueva York, Alfred A. Knopf, 2010, p. 4.

Capítulo 8: No calcules, analiza

1. Edgar Allan Poe, Los crímenes de la calle Morgue, Barcelona, Debate, reed. 2000.

2. Ibid.

3. Ibid.

Capítulo 9: El rompecabezas de Peirce (y Peirce se rompió la cabeza)

1. Joseph Brent, Charles Sanders Peirce: A Life, Bloomington, IN, Indiana

University Press, 1993, pp. 1–7.

2. Citado ibid. , pp. 2-3.

3. Henry James, ed., The Letters of William James, vol. 1, Boston, Atlantic Monthly Press, 1920, p. 35.

4. Citado en Brent, Peirce, p. 16.

5. Joseph Brent compara a Peirce con un dandi: «El dandi vive y duerme delante de un espejo; es rico y se deja consumir por su trabajo, que realiza de manera desinteresada. Es solitario e infeliz» (pp. 23-24). El novelista Henry James, hermano de William, comentó una vez en broma, tras encontrarse a Peirce desolado, deambulando por París justo después de la separación de su esposa, que este era interesante y vestía «ropa bonita» (p. 25). En efecto, la vida de Peirce fue motivo de una letanía de quejas por parte de sus seres queridos, colegas y superiores, quienes lo acusaron de ser ateo (no era cierto), alcohólico (quizá fuera cierto), adicto a las drogas (cierto, pero con motivo: durante toda su vida padeció un doloroso mal conocido como neuralgia facial), infiel en su matrimonio (sin duda) e imprudente en su desempeño profesional. Nunca se ha acabado de saber toda la verdad acerca de su despido como conferenciante en Hopkins. Su incapacidad para cuidar del costoso equipo con el que efectuaba mediciones gravitacionales mientras estaba a sueldo del Servicio de Costas está bien documentado, así como la lentitud con la que realizaba sus proyectos.

6. Brent, Peirce, p. 9.

7. Casi por definición, los descubrimientos científicos genuinos tienden a ser inferencias improbables (o constelaciones de inferencias) cuando se los juzga a partir de un «indicio» previo y aceptado. Cuando el descubrimiento se acepta también (en caso de que sea así), el indicio mismo se reevalúa, y se acompaña de juicios de probabilidad.

Capítulo 10: Problemas de deducción e inducción

1. Técnicamente, los sistemas universales de comprobación de errores son

imposibles, tal y como sabemos gracias a los teoremas de incompletitud de Gödel. Pero, en vista de esa salvedad, los sistemas deductivos resultan útiles para comprobar si el software se adecúa a una serie de especificaciones, entre otras cosas.

2. Wesley Salmon, Causality and Explanation, Oxford, Oxford University Press, 1998.

3. Adopto aquí una palabra habitual en la IA, «agentes», que significa cualquier cosa que actúe de manera independiente. Una persona puede ser un agente inteligente (o «cognitivo»), pero también pueden serlo un sistema de IA o un alienígena.

4. David Hume, Tratado de la naturaleza humana, Madrid, Tecnos, reed. 2018.

5. La formulación original de Russell hacía referencia a un pollo, no a un pavo. Fue Karl Popper quien reformuló el ejemplo bajo su forma actual. Y la cita está tomada de una fuente secundaria: Alan Chalmers, ¿Qué es esa cosa llamada ciencia?, Madrid, Siglo XXI, 2006.

6. Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Nueva York, Viking, 2019, p. 48.

7. Gary Marcus y Ernest Davis, Rebooting AI: Building Artificial Intelligence We Can Trust, Nueva York, Pantheon Books, 2019.

8. Ibid. , p. 62.

9. Martin Ford, Architects of Intelligence: The Truth about AI from the People Building It, Birmingham, UK, Packt Publishing, 2018.

10. Nassim Nicholas Taleb, «The Fourth Quadrant: A Map of the Limits of Statistics», Edge.org, 14 de septiembre de 2008. https://www.edge.org/conversation/nassim_nicholas_taleb-the-fourth-quadrant-a-map-of-the-limits-of-statistics.

11. Judea Pearl and Dana Mackenzie, The Book of Why: The New Science of Cause and Effect, Nueva York, Basic Books, 2018. (Trad. esp.: El libro del porqué, Barcelona, Pasado & Presente, 2020.)

12. Debería decir algo más sobre los términos «observación», «datos» y «hecho». Imagínate a Fred, un tipo que recibe el encargo de su jefe de recoger datos sobre todas las personas que lleven chaquetas North Face en una esquina concreta del centro de Chicago el 1 de diciembre entre las 7 de la mañana y las 7 de la tarde. Fred observa a los transeúntes, toma nota de las chaquetas que visten. Por cada chaqueta North Face que vea, registra esa observación en un bloc de notas. Cuando dan las 7 de la tarde, le lleva el bloc a su supervisor, quien se ha de encargar de tomar las observaciones registradas (el número 147) e introducirlas en una hoja de cálculo. Las observaciones de Fred se han convertido ahora en datos; han sido registradas en formato legible por un ordenador y representan el hecho del número de chaquetas North Face que han vestido los transeúntes en ese lugar y durante esas horas. Por consiguiente, los datos son observaciones registradas que se consideran hechos. En realidad, la cuestión al completo de lo que es un hecho resulta bastante interesante, pero tendrá que esperar a algún otro proyecto.

13. Al parecer, a Pearl también le interesa la labor detectivesca, ya que señala que Sherlock Holmes no usaba la deducción, según el héroe de ficción aseguraba a veces. En opinión de Pearl, se servía de la inducción, en el sentido de que examinaba los hechos para llegar a su explicación. Sin duda, Pearl no puede decirlo en sentido literal, ya que la afirmación implica contradecir su propio esquema de la escalera, pues la asociación de observaciones sin una comprensión difícilmente llevaría a solucionar los rompecabezas a los que se enfrenta Holmes. La combinación expone un malentendido en boga sobre la inducción en general: que las teorías surgen de una observación meticulosa. Esto es cierto solo si permitimos que una teoría previa conforme la observación y si, hablando en un sentido literal, la observación por sí sola no puede suministrar la teoría que necesitamos. Tal y como veremos con la comprensión del lenguaje, la inducción toma parte en una especie de círculo holístico hermenéutico donde las ideas previas pasan a ser interpretadas, y la interpretación en curso no deja de ajustar esas ideas previas además de sugerir ideas nuevas. Resulta imposible darle sentido al lenguaje basándose en una comprensión correcta de la inducción (esto es, sin agruparla junto a otras formas de inferencia).

14. Marcus y Davis, Rebooting AI.

15. Russell, Human Compatible.

16. Pearl, The Book of Why, p. 36.

Capítulo 11: El aprendizaje automático y los macrodatos

1. Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Nueva York, Viking, 2019, p. 48.

2. Tom Mitchell, Machine Learning, Nueva York, McGraw-Hill Education, 1997, p. 2.

3. En gran medida, nuestra confianza en los filtros se debe a que estos son deliberadamente permisivos: es más probable que llegue un correo basura a la bandeja de entrada a que se descarte un mensaje legítimo. Las consecuencias de usar sistemas que no ofrecen garantías y que carecen de una comprensión real son una fuente de preocupación cada vez mayor. El correo basura representa una ansiedad pequeña en comparación con los vehículos sin conductor.

4. Melanie Mitchell, Artificial Intelligence: A Guide for Thinking Humans, Nueva Yor,: Farrar, Straus, and Giroux, 2019.

5. Doug Laney, «3-D Data Management: Controlling Data Volume, Velocity and Variety», nota de investigación del Grupo Gartner, febrero de 2001. El original ya no se puede encontrar en Gartner, pero Laney lo ha vuelto a colgar en: https://community.aiim.org/blogs/doug-laney/2012/08/25/deja-vvvu-gartners-original-volume-velocity-variety-definition-of-big-data. Más tarde, IBM se apropió de ese esbozo tan pegadizo y le añadió una cuarta uve, la de «veracidad». Pero Laney señaló a continuación en una entrada de blog (a la que ya no se puede acceder online) que, en realidad, la veracidad es inversamente proporcional al tamaño, de modo que su añadido estropea la definición de Gartner. Laney explicó que aspectos como el de veracidad no son una medida de «grandura», y por tanto tampoco son un rasgo que defina los macrodatos más de una vez. Doug Laney, «Batman on Big Data», Garter Blog Network,13 de noviembre de 2013.

6. Gil Press, «12 Big Data Definitions: What’s Yours?»,Forbes, 3 de septiembre de 2014.

7. Véase: https://obamawhitehouse.archives.gov/blog/2016/05/23/administration-issues-strategic-plan-big-data-research-and-development.

8. Jonathan Stuart Ward y Adam Barker, «Undefined by Data: A Survey of Big Data Definitions», School of Computer Science, the University of Saint Andrews, UK. Publicado en Published in arXiv, 2013.

9. Viktor Mayer-Schönberger y Kenneth Cukier, Big Data: La revolución de los datos masivos, Madrid, Turner, 2013.

10. Chris Anderson, «The End of Theory: The Data Deluge Makes the Scientific Method Obsolete», Wired, 23 de junio de 2008.

11. Gil Press, «Big Data is Dead. Long Live Big Data AI», Forbes, 1 de julio de 2019.

12. Todas ellas se pueden registrar como respuestas a preguntas de tipo sí/no en una estructura de datos (llamada vector) que contenga «1» para SÍ y «0» para NO. Así {1, 1, 1, 0, 0, 1} es un vector de rasgos que se puede suministrar como entrada a un algoritmo de aprendizaje (el «aprendiz»). Hay muchas maneras de estructurar las entradas para los diferentes aprendices, esto es solo un ejemplo. Los detalles no importan tanto como la idea en general.

13. James Somers, «The Man Who Would Teach Machines To Think», The Atlantic, noviembre de 2013.

Capítulo 12: La inferencia abductiva

1. Charles Sanders Peirce, «Guessing», The Hound and Horn, 2, pp. 267–282, 271.

2. Ibid. , p. 271.

3. Ibid. , p. 272.

4. Ibid. , p. 277.

5. Gary Marcus y Ernest Davis, Rebooting AI: Building Artificial Intelligence We Can Trust, Nueva York, Pantheon, 2019, pp. 145–146.

6. Ibid. , p. 146.

7. Collected Papers of Charles Sanders Peirce, eds. Charles Hartshorne y Paul Weiss, vols. 1–6, Cambridge, MA, Harvard University Press, 1931–1935, p. 5189.

8. Charles Sanders Peirce Papers, Biblioteca Houghton, universidad de Harvard, manuscrito 692.

9. Melanie Mitchell, Artificial Intelligence: A Guide for Thinking Humans, Nueva York, Farrar, Straus, and Giroux, 2019.

10. Ibid.

11. Peirce Papers, manuscrito 692.

12. Collected Papers, p 5171.

13. El trabajo sobre la abducción experimentó un resurgimiento parcial en la década de 2010, en particular gracias a los intentos por convertirla en un tipo de inferencia probabilística o bayesiana. Esos enfoques han llamado la atención tanto de los especialistas como de quienes no lo son, cuando menos porque nos han ayudado a apreciar mejor el verdadero misterio de la abducción, entendida de manera correcta como una conjetura o hipótesis que explica un suceso u observación en el pensamiento cotidiano (y científico). Los sistemas bayesianos realizan siempre suposiciones simplistas, que por un lado consideran solo un pequeño subconjunto de posibles hipótesis conocidas (a veces apenas dos) y, por el otro, buscan inferencias que tengan una alta probabilidad, dado el ejemplo o caso en cuestión. Una vez más, esos sistemas, aunque se trate de extensiones y exploraciones interesantes del razonamiento probabilístico, llevan la abducción solo en el nombre. Para un buen resumen del resultado de las investigaciones sobre ese tema, véase el trabajo de Ray Mooney en la universidad de Texas en Austin: http://www.cs.utexas.edu/~ml/publications/area/65/abduction.

14. Hector Levesque, Common Sense, the Turing Test, and the Quest for Real

AI, Cambridge, MA, MIT Press, 2017.

15. Hector Levesque, «On Our Best Behavior», Artificial Intelligence 212, n.º 1 (2014): pp. 27–35.

16. Los proyectos de gran envergadura que han rastreado la red para seleccionar hechos y conocimiento en su interior también han resultado decepcionantes. Para un debate sobre el NELL (Never Ending Language Lerner) de Tom Mitchell, véase: Marcus y Davis, Rebooting AI, pp. 150–151.

17. John Haugeland ya señaló en 1979 la problemática de organizar el conocimiento para llegar a la inferencia: «El concepto de "mono" incluiría no solo que son un cierto tipo de primates [información taxonómica], sino un montón de información relacionada, como el lugar del que proceden, lo que comen, el uso que les dan los organilleros y lo que ese ejemplar tan grande que hay en el zoo les tira a los espectadores». Haugeland se preguntó cómo usar la información «típica» sobre el concepto de mono cuando no hay nada en la información enciclopédica que sea relevante. Y la cuestión sigue siendo válida. Véase: John Haugeland, «Understanding Natural Language», The Journal of Philosophy 76, n.º 11 (1979), p. 623

18. R. C. Schank, Conceptual Information Processing, Nueva York, Elsevier, 1975.

19. Peirce, «Guessing», p. 269.

20. Ibid.

21. Ibid.

22. Kahneman, Pensar rápido, pensar despacio.

23. Umberto Eco y Thomas A. Sebeok (eds.), El signo de los tres (Dupin, Holmes, Peirce), Barcelona, Lumen, 1989.

Capítulo 13: Inferencia y lenguaje 1

1. Andrew Griffin, «Turing Test Breakthrough as Super-Computer Becomes First to Convince Us It’s Human», Independent, 8 de junio de 2014.

2. Véase:«Computer AI Passes Turing Test in ‘World First’», BBC News,9 de junio de 2014. El artículo de Time ya no se puede recuperar. Véase también: Pranav Dixit, «A Computer Program Has Passed the Turing Test for the First Time», Gizmodo, 8 de junio de 2014.

3. Gary Marcus, «What Comes After The Turing Test?», New Yorker,9 de junio de 2014.

4. Adam Mann, «That Computer Actually Got an F on the Turing Test», Wired, 9 de junio de 2014.

5. Podríamos convertir a Siri o a Cortana en contendientes para el premio Loebner añadiéndoles algún tipo de código como hazBalbucear() o hazQuejar(), donde los argumentos fueran órdenes o preguntas del usuario. A continuación, el sistema se quejaría, de manera fiable, sobre las peticiones humanas —pongamos que preguntando siempre «por qué», para lamentarse acto seguido como un invitado incómodo por lo cansado que está de tener que responder siempre a lo mismo, o exigiendo hablar sobre algún otro tema—. Así, en general, Siri podría parecer considerada e inteligente, a la vez que se volvería inútil por completo. Siri se volverá más útil, precisamente, cuando llegue al punto de comprender el lenguaje humano, que es uno de los motivos por los que el test de Turing aún no se ha superado, y quizá también por el que los científicos de la IA parecen tener tantas ganas de desecharlo.

6. Stuart Russell, Human Compatible: Artificial Intelligence and the Problem of Control, Nueva York, Viking, 2019.

7. Martin Ford, Architects of Intelligence: The Truth about AI from the People Building It, Birmingham, UK, Packt Publishing, 2018.

8. Gary Marcus y Ernest Davis, Rebooting AI: Building Artificial Intelligence We Can Trust, Nueva York, Pantheon Books, 2019, pp. 6–7.

9. Véase: Gary Marcus, «Why Can’t My Computer Understand Me?», New Yorker, 14 de agosto de 2013. El artículo de Levesque que se menciona en él se puede encontrar aquí: https://www.cs.toronto.edu/~hector/Papers/ijcai-13-paper.pdf. El enfoque de Marcus sobre el trabajo de Levesque resulta

esclarecedor y vale la pena leerlo. El artículo de Levesque para el IJCAI es apto para neófitos en la materia y ofrece un gran resumen del problema de los macrodatos para la comprensión del lenguaje. El enfoque de Marcus sobre el trabajo de Levesque resulta esclarecedor y vale la pena leerlo. El artículo de Levesque para el IJCAI es apto para neófitos en la materia y ofrece un gran resumen del problema de los macrodatos para la comprensión del lenguaje.

10. El último Desafío de Esquemas de Winograd que dispuso de patrocinio se celebró durante la Conferencia Internacional Conjunta sobre Inteligencia Artificial de 2016. El sistema ganador obtuvo un 58,3 % de acierto en el conjunto de datos, porcentaje insuficiente para optar al premio. Desde entonces, el patrocinador ha rechazado cubrir la bolsa de 25.000 dólares.

11. Hector Levesque, «On Our Best Behavior». Charla, Conferencia Internacional Conjunta sobre Inteligencia Artificial de Pekín, China, en 2013 (p. 4). https://www.cs.toronto.edu/~hector/Papers/ijcai-13-paper.pdf.

Capítulo 14: Inferencia y lenguaje 2

1. El análisis pragmático es el Santo Grial de las grandes compañías tecnológicas. A Google le encantaría insertar anuncios sobre vuelos baratos a Mazatlán, o quizá sobre paraguas baratos, en una página web al lado de un comentario que dijera «¡¡¡Aquí hace un tiempo fabuloso!!!» pero que se hubiera realizado en Seattle en diciembre (cuando lo más probable es que estuviera lloviendo e hiciera frío). Por desgracia, los anuncios que con mayor seguridad acabarían insertados allí serían de crema con protección solar 50 y bebidas con sombrillita. A la gente no le cuesta (en general) detectar el sarcasmo, pero este se queda fuera del reino de la IA contemporánea.

2. A. M. Turing, «Computing Machinery and Intelligence», Mind 59, n.º 236 (octubre de 1950), p. 446.

3. Este ejemplo subraya también el uso contextual de la preposición «on» en inglés. Así, «on» se utiliza para decir que uno está en un barco, pero en el caso de los coches se usa solo «in». De igual manera, la mayoría de la gente prefiere

«on» antes que «in» para los trenes. Los aviones son complicados: se utiliza «on» para hablar del vuelo e «in» para referirse al aparato. Pero todos estos ejemplos se pueden capturar a partir de los patrones de uso, ya que son expresiones que se encuentran generalmente en los textos. Los problemas surgen cuando aparece un uso atípico: «I'm on the house»... literalmente, «estoy sentado encima de la casa», que es un ejemplo de elisión, de dejar cosas fuera porque se entienden a través del contexto.

4. Un equipo de Google Brain obtuvo un 61,5 % de acierto en una prueba difícil de esquemas de Winograd que se conoce como WSC-273. El sistema, de última generación, consiguió aumentar en un 10 % la efectividad respecto a las respuestas aleatorias, lo cual demuestra antes el carácter irresoluble de la prueba cuando se usan métodos de datos que cualquier avance legítimo. Ten en cuenta también que el equipo tuvo acceso al conjunto de datos que se usó durante las fases de entrenamiento y de comprobación, y que analizó las preguntas de manera extensiva —a duras penas se trataría de una prueba de capacidad lingüística que cualquier persona pudiera realizar se supone que felizmente y de manera espontánea.

5. Paul Grice, Studies in the Way of Words, Cambridge, MA, Harvard University Press, 1991.

6. Levesque, «On Our Best Behavior», p. 1.

7. En las pruebas de esquemas de Winograd, algunas de las preguntas en las que yerran los sistemas de IA tienden a ser parecidas a las que se saldan con aciertos, y no son más difíciles que estas. Ese escenario también es negativo, porque no hay manera de explicar los errores. Es solo que el sistema se equivoca con algunas preguntas que a todos los efectos parecen ser exactamente iguales que aquellas en las que acierta, siendo la única diferencia que, al escogerlas al azar, el sistema no ha entendido nada desde un principio. El debate reciente sobre la posibilidad de confiar en los sistemas de la IA contemporánea cuando no entendemos por qué se equivocan (o aciertan) apuntan al corazón de una cuestión cada vez más perturbadora, a medida que la IA basada en datos se vuelve más generalizada.

8. D. A. Ferrucci, «Introduction to "This is Watson"», IBM Journal of Research and Development 56, n.º 3.4 (3 de abril de 2012), pp. 1:1–1:15.

9. Marvin Minsky, The Society of Mind, Nueva York, Simon & Schuster, 1988.

10. J. Fan, A. Kalyanpur, D. C. Gondek, D. A. Ferrucci, «Automatic Knowledge Extraction from Documents», IBM Journal of Research and Development 56, n.º 3-4 (2012), pp. 5:1–5:10.

11. Marcus y Davis, Rebooting AI, p. 14.

12. Marcus y Davis discuten las diferencias entre las preguntas de trivialidades y de «sentido común» en la sección sobre la comprensión del lenguaje y la IA. Véase, por ejemplo: ibid. , pp. 74-76.

13. Citado ibid. , p. 68.

14. Ibid. , p. 69.

15. Ibid. , p. 27.

16. Charles Sanders Peirce, «Logical Machines», American Journal of Psychology 1, n.º 1 (1887): p. 165.

17. Ibid. , p. 168.

18. Ibid. , p. 169.

19. Ibid. , p. 169.

Capítulo 15: Mitos y héroes

1. Jaron Lanier, You Are Not a Gadget: A Manifesto, Nueva York, Alfred A. Knopf, 2010, p. 2.

2. Clay Shirky, Excedente cognitivo. Creatividad y generosidad en la era conectada, Barcelona, Deusto, 2012.

3. Yochai Benkler, La riqueza de las redes. Cómo la producción social

transforma los mercados y la libertad, Barcelona, Icaria, 2015.

4. Ibid. , epígrafe.

5. Clay Shirky, Here Comes Everybody: The Power of Organizing Without Organizations, Nueva York, Penguin Books, 2009.

6. Lanier, You Are Not a Gadget, p. 1.

7. James Surowiecki, Cien mejor que uno. La sabiduría de la multitud o por qué la mayoría es más inteligente que la minoría, Barcelona, Urano, 2005.

Capítulo 16: La mitología de la IA invade la neurociencia

1. Sean Hill, «Simulating the Brain», en Gary Marcus y Jeremy Freeman, eds., The Future of the Brain: Essays by the World’s Leading Neuroscientists, Princeton, NJ, Princeton University Press, 2015, pp. 123–124

2. Henry Markram, «Seven Challenges for Neuroscience», Functional Neurology 28 (2013), pp. 145–151.

3. Ed Yong, «The Human Brain Project Hasn’t Lived Up to Its Promise», The Atlantic, 22 de julio de 2019.

4. The Kavli Foundation, «The BRAIN Initiative: Surviving the Data Deluge», https://www.kavlifoundation.org/science-spotlights/brain-initiative-surviving-data-deluge#.XgVezkdKhdg.

5. Amye Kenall, «Building the Brain: The Human Brain Project and the New Supercomputer», BioMed Central, 8 de julio de 2014. http://blogs.biomedcentral.com/bmcblog/2014/07/08/building-the-brain-the-human-brain-project-and-the-new-supercomputer/.

6. Rebecca Golden, «Mind-Boggling Numbers: Genetic Expression in the Human Brain», Science 2.0, 15 de abril de 2013. https://www.science20.com/rebecca_goldin/mindboggling_numbers_genetic_exp

109345

7. Yves Frégnac y Gilles Laurent, «Neuroscience: Where Is the Brain in the Human Brain Project?», Nature, 3 de septiembre de 2014.

8. Ibid.

9. Ibid.

10. Eric Kandel, Henry Markram, Paul M. Matthews, Rafael Yuste y Christof Koch, «Neuroscience Thinks Big (and Collaboratively)», Nature Reviews Neuroscience 14, n.º 9 (2013), p. 659.

11. Lee Gomes, «Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts», IEEE Spectrum, 20 de octubre de 2014. https://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts

12. Ibid.

13. Gary Marcus y Ernest Davis, «Eight (No Nine!) Problems with Big Data», New York Times, 6 de abril de 2014.

14. Nate Silver, La señal y el ruido. Cómo navegar por la maraña de datos que nos inunda, localizar los que son relevantes y utilizarlos para elaborar predicciones infalibles, Barcelona, Península, 2014.

Capítulo 17: Las teorías de la inteligencia humana basadas en el neocórtex

1. Jeff Hawkins, Sobre la inteligencia, Barcelona, Espasa, 2005.

2. Ray Kurzweil, How to Create a Mind: The Secret of Human Thought Revealed, Nueva York, Penguin Books, 2013, p. 35.

3. Gary Marcus, «Ray Kurzweil's Dubious New Theory of Mind», New Yorker, 15 de noviembre de 2012.

4. Ibid.

5. Para una descripción general, véase: Ferris Jabr, «Memory May be Built with Standard Building Blocks», New Scientist, 17 de marzo de 2011.

6. Henry Markram y Rodrigo Perin, «Innate Neural Assemblies for Lego Memory», Frontiers in Neural Circuits 5 (2011), p. 6.

7. Eric Kandel, Henry Markram, Paul M. Matthews, Rafael Yuste y Christof Koch, «Neuroscience Thinks Big (and Collaboratively)», Nature Reviews Neuroscience 14, n.º 9 (2013), p. 659.

8. Yves Frégnac y Gilles Laurent, «Neuroscience: Where Is the Brain in the Human Brain Project?», Nature News 513, n.º 7516 (2014), p. 27

9. «Carta abierta a la Comisión Europea en relación al Proyecto Cerebro Humano». Véase: https://neurofuture.eu/.

10. Eliza Shapiro, «Obama Launches BRAIN Initiative to Map the Human Brain», Daily Beast, 3 de abril de 2013. Actualizado el 11 de julio de 2017.

11. Véase: Stefan Thiel, «Why the Human Brain Project Went Wrong—and How to Fix It», Scientific American, 1 de octubre de 2015. También Ed Yong, «The Human Brain Project Hasn't Lived Up to Its Promise», The Atlantic, 22 de julio de 2019.

12. Yong, «The Human Brain Project Hasn't Lived Up».

Capítulo 18: ¿El fin de la ciencia?

1. Peter Thiel, entrevistado por Eric Weinstein en el podcast The Portal, episodio #001: «An Era of Stagnation & Universal Institutional Failure», 19 de julio de 2019. https://www.youtube.com/watch?v=nM9f0W2KD5s&t=1216s.

2. Ibid.

3. Norbert Wiener, Invention: The Care and Feeding of Ideas, Cambridge, MA, MIT Press, 1994.

4. Ibid. , p. 89.

5. Ibid. , p. 96.

6. Ibid. , p. 96.

7. Ibid. , p. 87.

8. Ibid. , p. 87.

9. John Horgan, El fin de la ciencia. Los límites del conocimiento en el declive de la era científica, Barcelona, Paidós, 1998.

Notas del traductor

[*a volver] En la traducción de Rafael Cansinos Assens (N. del T.).

[*b volver] «La caja», en español (N. del T.).

[*c volver] El «window» del original es «ventana» y «escaparate» (N. del T.).

[*d volver] El billón norteamericano se refiere por lo general a mil millones (N. del T.).

[*e volver] Recuérdese la ambigüedad ya comentada entre «pen» como «bolígrafo» y como «jaula» o «parque infantil» (N. del T.).

[*f volver] En traducción de Manuel Mujica Lainez (N. del T.).

[*g volver] El pentámetro yámbico es un verso en el que se suceden cinco pares de sílabas con acento en la segunda sílaba de cada par. Ante la imposibilidad de

replicar ese ritmo con el nombre en español de las estaciones, nos contentamos con la equiparación entre las tres sílabas de «verano» e «invierno», respecto a las cuatro de «primavera» (N. del T.).

[*h volver] «Brain» significa «cerebro» (N. del T.).

[*i volver] En el original, «brave new world», título original de la obra distópica Un mundo feliz, que Aldous Huxley tomó de un verso de La tempestad de Shakespeare (N. del T.).